International Journal of Computer Science & Information Security

Editorial
Message from Managing Editor

We thank all those authors who contributed papers to the May 2010 issue and the reviewers, all of whom provided valuable feedback comments. We hope that you will find this IJCSIS edition a useful state-of-the-art literature reference for your research projects.

We look forward to receiving your submissions and to receiving feedback.

IJCSIS May 2010 Issue (Vol. 8, No. 2) has an acceptance rate of 26%.

Special thanks to our technical sponsors for their valuable service.

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

18. Paper 20041030: An Efficient OFDM Transceiver Design Suitable to IEEE 802.11a WLAN standard (pp. 118-122)

T. Suresh, Research Scholar, R.M.K Engineering College, Anna University, Chennai TamilNadu, India
Dr. K. L. Shunmugathan, Professor & Head, Department of CSE, R.M.K Engineering College, Kavaraipettai, TamilNadu, India

19. Paper 20041031: Comparitive Analysis of Smart Antenna Array, Basis of Beamforming Schemes and Algorithems : A Review (pp. 123-128)

Abhishek Rawat , R. N. Yadav and S. C. Shrivastava
Maulana Azad National Institute Of Technology, Bhopal, INDIA

20. Paper 20041032: Comments on Five Smart Card based Password Authentication Protocols (pp. 129-132)

Yalin Chen [1], Jue-Sam Chou [2,] , Chun-Hui Huang [3]*
[1] Institute of information systems and applications, National Tsing Hua University, Taiwan
[2] Department of Information Management, Nanhua University, Taiwan
[3] Department of Information Management, Nanhua University, Taiwan

21. Paper 20041033: Cryptanalysis on Four Two-party Authentication Protocols (pp. 133-137)

Yalin Chen [1], Jue-Sam Chou [2,] , Chun-Hui Huang [3]*
[1] Institute of information systems and applications, National Tsing Hua University, Taiwan
[2] Department of Information Management, Nanhua University, Taiwan
[3] Department of Information Management, Nanhua University, Taiwan

22. Paper 20041034: Software Metrics: Some Degree of Software Measurement and Analysis (pp. 138-144)

Rakesh. L, Department of Computer-Science, SCT Institute of Technology, Bangalore, India-560075
Dr. Manoranjan Kumar Singh, PG Department of Mathematics, Magadh University, Bodhagaya, India-824234
Dr. Gunaseelan Devaraj, Department of Information Technology, Ibri college of Technology, Ibri, Sultanate of Oman- 516

23. Paper 22041038: Preprocessing of Video Image with Unconstrained Background for Drowsy Driver Detection (pp. 145-151)

M. Moorthi [1], Dr. M.Arthanari [2], M.Sivakumar [3]
[1] Assistant Professor, Kongu Arts and Science College, Erode – 638 107, Tamil Nadu, India
[2] Prof. & Head, Tejaa Sakthi Institute of Technology for Women, Coimbatore – 641 659, Tamil Nadu, India
[3] Doctoral Research Scholar, Anna University, Coimbatore, Tamil Nadu, India

24. Paper 23041041: Ultra Fast Computing Using Photonic Crystal Based Logic Gates (pp. 152-155)

X. Susan Christina, Dept. of ECE, Mookambigai College of Engg., Trichy- 622 502, India.
A. P. Kapilan, Dept. of ECE, Chettinad College of Engg & Tech, Karur,. 639114. India.
P. Elizabeth Caroline, Dept. of ECE, JJ College of Engg &Tech, Trichy –620 009,India

25. Paper 25041043: Markov Chain Simulation of HIV/AIDS Movement Pattern (pp. 156-167)

Ruth Stephen Bature, Department of Computer/Mathematical Science, School of Science Technology, Federal College of Chemical and Leather Technology, Zaria, Nigeria.
Obiniyi, A. A., Department of Mathematics, Ahmadu Bello University, Zaria, Nigeria
Ezugwu El-Shamir Absalom, Department of Mathematics, Ahmadu Bello University, Zaria, Nigeria
Sule, O. O., Department of Computer/Mathematical Science, School of Science Technology, Federal College of Chemical and Leather Technology, Zaria, Nigeria.

35. Paper 28041059: Effective MSE Optimization in Fractal Image Compression (pp. 238-243)

A. Muruganandham, Sona College of Technology, Salem-05.,India.
Dr. R. S. D. Wahida banu, Govt Engineering College, Salem-11, India

36. Paper 28041061: Embedding Expert Knowledge to Hybrid Bio-Inspired Techniques- An Adaptive Strategy Towards Focussed Land Cover Feature Extraction (pp. 244-253)

Lavika Goel , M.E. (Masters) Student , Delhi College of Engineering, New Delhi, India
Dr. V.K. Panchal, ADD.DIRECTOR & SCIENTIST 'G', Defence Terrain & Research Lab(DTRL) , DRDO , Delhi
Dr. Daya Gupta, HEAD OF DEPARTMENT, Computer Engineering Department, Delhi College of Engineering, Delhi

37. Paper 29041063: On Multi-Classifier Systems for Network Anomaly Detection and Features Selection (pp. 254-263)

Munif M. Jazzer, Faculty of ITC, Arab Open University-Kuwait, Kuwait.
Mahmoud Jazzar, Dept. of Computer Science, Birzeit University, Birzeit, Palestine
Aman Jantan, School of Computer Sciences, University of Science Malaysia, Pulau Pinang, Malaysia

38. Paper 29041066: AccSearch: A Specialized Search Engine for Traffic Analysis (pp. 264-271)

K. Renganathan, Computer Science and Engineering Department, SRM University, India
B. Amutha, Computer Science and Engineering Department, SRM University, India

39. Paper 30031085: A Study of Voice over Internet Protocol (pp. 272-278)

Mohsen Gerami, The Faculty of Applied Science of Post and Communications, Danesh Blv, Jenah Ave, Azadi Sqr, Tehran, Iran.

40. Paper 30041067: Performance Issues of Health Care System using SQL Server (pp. 279-284)

Narendra Kohli, Electrical Engineering Department, Indian Institute of Technology, Kanpur, India
Nishchal K. Verma, Electrical Engineering Department, Indian Institute of Technology, Kanpur, India

41. Paper 30041068: Color Steel Plates Defect Detection Using Wavelet And Color Analysis (pp. 285-292)

Ebrahim Abouei Mehrizi, Department of Electronic Engineering, Islamic Azad University, najafabad branch, Isfahan, 81746, Iran
Amirhassan Monadjemi, Department of Computer Engineering, University of Isfahan, Isfahan, 81746, Iran
Mohsen Ashorian, Department of Electronic Engineering, Islamic Azad University, shahremajlesi branch, Isfahan, 81746, Iran

42. Paper 30041069: Clustering in Mobile Ad hoc Networks: A Review (pp. 293-301)

Meenu Chawla, Department of CSE, MANIT, Bhopal, India
Jyoti Singhai, Department of ECE, MANIT, Bhopal, India
J L Rana, Department of CSE, MANIT, Bhopal, India

43. Paper 30041071: Survey of Nearest Neighbor Techniques (pp. 302-305)

Nitin Bhatia, Department of Computer Science, DAV College, Jalandhar, India
Vandana, SSCS, Deputy Commissioner's Office, Jalandhar, India

44. Paper 30041081: Time Domain Analysis Based Fault Diagnosis Methodology for Analog Circuits - A Comparative Study of Fuzzy and Neural Classifier Performance (pp. 306-313)

V. Prasannamoorthy 1, R. Bharat Ram 2, V. Manikandan 3, N. Devarajan 4
1, 2, 4 Department of Electrical Engineering, Government College of Technology Coimbatore, India
3 Department of Electrical Engineering, Coimbatore Institute of Technology Coimbatore, India

45. Paper 30041082: Evaluation of English-Telugu and English-Tamil Cross Language Information Retrieval System using Dictionary Based Query Translation Method (pp. 314-319)

P. Sujatha , Department of Computer Science, Pondicherry Central University, Pondicherry-605014, India
P. Dhavachelvan, Department of Computer Science, Pondicherry Central University, Pondicherry-605014, India
V. Narasimhulu, Department of Computer Science, Pondicherry Central University, Pondicherry-605014, India

46. Paper 30041086: A Novel Approach for Hand Analysis Using Image Processing Techniques (pp. 320-323)

Vishwaratana Nigam, Divakar Yadav, Manish K Thakur
Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India

47. Paper 30041087: Applying l-Diversity in Anonymizing Collaborative Social Network (pp. 324-329)

G. K. Panda, Department of CSE & IT, MITS, Sriram Vihar Rayagada,, India
A. Mitra, Department of CSE & IT , MITS, Sriram Vihar Rayagada,, India
Ajay Prasad, Department of CSE, Sir Padampat Singhania University, Udaipur, India
Arjun Singh, Department of CSE, Sir Padampat Singhania University, Udaipur, India
Deepak Tour, Department of CSE, Sir Padampat Singhania University, Udaipur, India

48. Paper 30041072: 3D-Mesh Denoising Using an Improved Vertex based Anisotropic Diffusion (pp. 330-337)

Mohammed El Hassouni, DESTEC, FLSHR, University of Mohammed V-Agdal- Rabat, Morocco
Driss Aboutajdine, LRIT, UA CNRST, FSR, University of Mohammed V-Agdal-Rabat, Morocco

49. Paper 30041083: A New Approach for Security Risk Assessment Caused by Vulnerabilities of System by Considering the Dependencies (pp. 338-346)

Mohammad Taromi, Performance and Dependability Eng. Lab., School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
Mohammad Abdollahi Azgomi, Performance and Dependability Eng. Lab., School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

50. Paper 28041060: Image Super Resolution Using Marginal Ditribution Prior (pp. 347-351)

S. Ravishankar, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham University Bangalore, India
Dr. K.V.V. Murthy, Department of Electronics and Communication, Amrita Vishwa Vidyapeetham University, Bangalore, India

51. Paper 10041012: A Survey on WiMAX (pp. 352-357)

Mohsen Gerami, The Faculty of Applied Science of Post and Communications, Danesh Blv, Jenah Ave, Azadi Sqr, Tehran, Iran.

Policy-based Self-Adaptive Media Service Architecture for Reliable Multimedia Service Provisioning

[1]G. Maria kalavathy, [2]N. Edison Rathinam and [3]P. Seethalakshmi
[1]Sathyabama University, Chennai, India
[2]Madras university Chennai, India
[3]Anna university Tiruchirapalli, India

Abstract— **The main objective of this paper is to design and develop the Self-Adaptive Media Service Architecture (SAMSA) for providing reliable multimedia services through policy-based actions. The distributed multimedia services deployed using SOA can be accessed in heterogeneous environments that are prone to changes during run-time. To provide reliable multimedia services, a powerful self-adaptable architecture is necessary to adapt at run time and react to the environment. The adaptability in this proposed architecture is achieved by enabling the service providers to Monitor, Analyze and Act on the defined policies that support customization of compositions of multimedia services. The Media Service Monitor (MSM) observes the business and quality metrics associated with the media services at run-time. The Adaptive Media Service Manager (AMSM) takes corrective actions based on the monitored results, through the policies defined as an extension of WS-Policy. The effectiveness of the proposed SAMSA has been evaluated on Dynamic Composite Real-Time Video on Demand Web Service (DCRVWS) for a maximum of 200 simultaneous client's requests. The analysis of results shows that the proposed architecture provides 20% improvement on reliability, response time and user satisfaction.**

Index Terms— **DCRVWS, Media Service Monitor, Reliable Multimedia Service, SAMSA**.

I. INTRODUCTION

Emerging advances in distributed multimedia services, such as video conferencing, media-on-demand and multimedia streaming demands scalable, robust and adaptive architecture for providing better reliable multimedia services. The adaptive architecture enables the flexible composition of multimedia services and improves the non-functional requirements of the multimedia services. The quality requirements of multimedia services and the expectations of end-users regarding the perceived service quality is becoming a major concern for multimedia service providers. The guaranteed quality service provisioning in case of failures, composing reliable multimedia web services incorporating run-time changes are challenging tasks that are to be addressed. These challenges are addressed at various layers such as service provider layer, transport layer, SOAP messaging layer, and business process layer. This paper addresses the reliability in service provider layer by deploying the load balanced redundant multimedia web services, in SOAP Messaging layer by monitoring SOAP messages and in business process layer by customization of activities at run time. To provide reliable and adaptive multimedia services, the powerful self-adaptable architecture is necessary which modifies its own behavior in response to changes in the observable Non-Functional Requirements (NFR) of the multimedia services. The proposed SAMSA utilizes the basic principles of SOA [1] but its service provider changes its capabilities of providing services at runtime when the performance degrades. The service provider of SAMSA includes run-time components such as monitor, analyzer, and corrective action taker to enable self-adaptability. The monitor component senses the run-time performance of the multimedia services such as response time, external errors and percentage of successful completion of multimedia web services using Parallel Performance Monitoring Service (PPMS) [2]. The monitored results are analyzed by analyzer for categorizing the type of faults such as timeout, user interruption, service failure, service unavailability, SLA violation, web server overload and network fault. According to the monitored results and type of fault, the adaptation has to be done at run-time based on the adaptation policies.

This paper focuses on building reliable composite multimedia web service with autonomous behavior capabilities such as self-healing and self-configuring [3] using SAMSA based on the adaptation policies. The adaptation can be done in different ways such as customization, correction, optimization and prevention. The customization demands addition, removal/replacement of components and its composition at run-time. The correction technique handles the faults detected during execution of the component. The optimization improves non-functional issues of services. The prevention mechanism prevents future faults or non-functional issues before its occurrence. This classification of adaptations is similar to the classification of software evolution into adaptive, corrective, perfective and preventive [4]. This paper focuses on the self-adaptable architecture with customization of composition and correction of failures that are detected during run-time.

This paper is organized as follows: Section 2 compares the

proposed architecture with other related works, the proposed Policy-based Self-Adaptive Media Service Architecture (SAMSA) is described in section 3 with its components, the case study DCRVWS is described in section 4, the Policy-based SAMSA is evaluated on the case study in section 5, and the last section summarizes the conclusion.

II. RELATED WORK

The dynamic media web service composition concepts are prominent approaches to advance construction of large scale distributed media services in a scalable, easy-programmable and efficient manner. According to user preferences in terms of QoS parameters the media service composition can be made flexible. The QoS-based web service selection and composition in service-oriented applications has gained more attention of many researchers [5][6]. Several ongoing academic and industrial efforts recognize the need to extend dynamic web service composition middleware with corrective adaptation to increase the reliability. This paper has unique characteristics of SAMSA to build reliable composite media web service through customization of composition at run time using policy-based approach. Because service-based software development for multimedia applications is emerging technology, there have been no reported performance assurance studies on multimedia web services. Most of the performance assurance testing is performed before the deployment of the web services. The MSU video quality measuring tool [7] measures video quality using metrics such as peak-to-peak signal-to-noise ratio (PSNR), Delta, MSAD (mean absolute difference of the color components) , MSE SSIM Index (measuring of three components luminance similarity, contrast similarity and structural similarity), VQM (uses DCT to correspond to human perception), MSU Blurring and MSU Blocking. These metrics are used in standalone applications and not used to measure the run-time performance of multimedia web services. Liguo Yu [8] proposed software wrapping technique at client side and the clients interact with the service through the wrapper which customize the messages exchanged between client and service and monitors the performance of the service by calculating the response time only. But in this paper, the media service monitor calculate response time using software wrapping technique from the service provider point of view to take immediate action if the performance is poor. Khaled Mahbub et al. [9] described the framework to monitor behavioral properties and assumptions at run time using event calculus. William N. Robinson [10] proposed REQMON monitoring system that raises only an alert by sending a failure message to the global monitor. Arne Koschel and Irina Astrova [11] designed a configurable event monitoring web service which is useful in the context of Event Driven Architectures (EDA) and Complex Event Processing (CEP). Onyeka Ezenwoye and S.Masoud Sadjadi [12] presented an approach to transparently adapting BPEL processes to tolerate runtime and unexpected faults and to improve the performance of overly loaded web services. They presented an another approach in which when one or more partner services do not provide satisfactory service the request for service is redirected to one of these

static, dynamic and generic proxies, where the failed or slow services are replaced by substitute services [13] and is not used for recovering the web service from the point at which the fault is occurred.

The proposed work differs from some recently published works for improving reliability of web service composition are discussed as follows: the RobustBPEL presented in [14] increases the reliability of BPEL processes through automatic generation of exception handling BPEL constructs, as well as generation of web service proxy to discover and bind to equivalent web service that can substitute a faulty service. Our work controls dynamic composition through customization using policies that can be checked for consistency. The aspect-oriented extension to BPEL was suggested in [15] to enable dynamic weaving of aspects into web service compositions. The QoS aspects they tried to address are security and state persistence which can be addressed at low-layer messaging middleware. The enforcement of quality using policies in our approach can be either delegated to SOAP messaging layer that mediates the web services interaction or enacted by BPEL engine through corrective adaptation. The service monitoring approach presented in [16] uses Web Service Constraint Language (WS-CoL) for specifying client-side monitoring policies that are related to security. The monitoring policies are specified external to process specification and achieves the desired reusability and separation of concerns. But it only provides support for monitoring and focuses mainly on security. Our work focuses on customization of processes at run-time and handling faults and address undesirable situations. The work in [17] proposed a general extension of SOA to support autonomic behavior of web services, but the proposed architecture does not address the requirements of self-adaptive business process execution. The task based recovery policies advocated by [18] are part of extended Petri net model, called Self-Adaptive Recovery Net (SARN), for specifying exceptional behavior in Petri net based workflow systems. The SARN recovery constructs are tightly coupled with Petri net concepts such as places, transitions and tokens. But the proposed adaptive policies are generic construct that model the required modifications to adapt the business process when a task failure event occurs. The policy-based approach is built on emerging self-healing architectures presented in [19].

III. POLICY-BASED SELF-ADAPTIVE MEDIA SERVICE ARCHITECTURE

The Policy-based SAMSA utilizes the basic principles of SOA, but provides dynamic services at runtime based on policies in response to operating environment that are implemented as the extension of WS-Policy [20]. The Policy-based SAMSA for multimedia applications includes the components such as User Profile Manager, User Preference Gatherer, SLA generator, Composition engine, and Adaptive Media Service Manager (AMSM) as shown in the figure 1. The Adaptive Media Service Manager comprises of the run-time components such as Media Service Monitor (MSM), Monitored Results Analyzer (MRA), Adaptation Policy Repository (APR), Adaptation Policy Parser (APP), QoS Renegotiator (QR) or SLA Regenerator and Load Balancer (LB). The Composition engine includes Web Service Map

(WSM) and Media Content Adapter (MCA). The Adaptive Media Service Manager monitors the non-functional requirements of media services using MSM and if any unforeseen behaviours are found, the faults are analyzed using MRA. Based on the results of MRA, the Adaptive Media Service Manager (AMSM) parses the Adaptation Policy Repository using Adaptation Policy Parser (APP) to take corrective action. The AMSM performs renegotiation (regeneration of SLA) with the user when the performance of service degrades and balances the load among the multiple Web Service Hosts (WSHs).
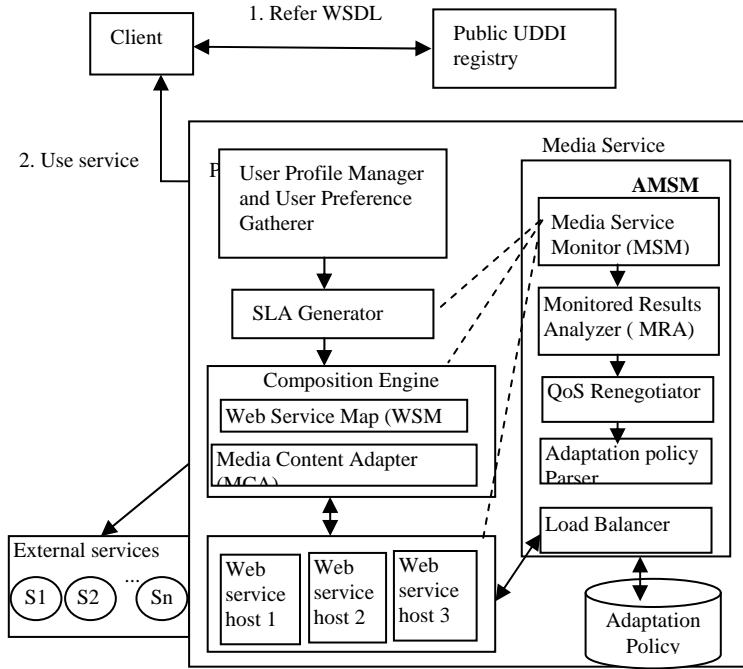
Fig. 1. Policy-based SAMSA for multimedia applications

A. User Profile Manager and User Preference Gatherer

The User Profile Manager component supports the candidate services such as authentication and new user registration. The user information such as name, location and field of interest are collected during registration for multimedia services. When the authenticated user is accessing the service next time, it displays the information about the history of services used. For example, if any VoD service is not fully viewed, it provides the details about the current status of the video that is recently viewed. During the first time accessing of service, the requirements for multimedia services and the expectations of end-users regarding the perceived service quality are collected using User Preference Gatherer. It displays the available services with its quality and cost and allows the user to select among them. When the requested quality of the service is not available, then the negotiation process is automated with the user through SLA generator or the quality of media content is transcoded using Media Content Adapter.

B. SLA Generator

SLA Generator has been designed to allow the automatic negotiation of QoS aspects between the user and service provider. It is implemented as a tuning service that provides

an user with a Graphical User Interface (GUI) to get the requirements such as perceptual quality and cost. These user requirements are mapped into QoS parameters such as frame rate and frame resolution. The user and provider can negotiate the maximum QoS possible within the budget constraint of the user using bilateral negotiation. In the bilateral negotiation, the service provider is not allowed to modify the QoS value proposed by the service user. Only the service user can modify the requested QoS value and suggest a lower bound value which is in the acceptable range of the application. For example, the frame resolution of the video clip requested by the service user is 320x240 pixels denoted as frreq, but the service provider offers the same video clip with the frame resolution of 328X208 / 720X576. Assuming that no content adaptation for this request is available and the user accepts 328x208 resolution denoted as frconfirm>=frreq, and then it is provided to the user. If the media content adaptation services such scaling, transcoding and bit-depth reduction are available, then the content is adapted according to the requirements of the service user. Automating the negotiation at design-time not only saves time and but also simplifies the run-time phase optimization. The negotiated QoS profiles are stored in the form of XML. Depending on the SLA generated, the candidate services are discovered from the available WSHs or from external service providers.

C. Web Service Map

The service retrieval process is done by searching the Web Service Map (WSM) using parallel search. The Parallel search algorithm proposed by Khitrin et al [21], was implemented using spatial encoding [22] has been used to search the WSM. The NOT-Shift-AND parallel search algorithm described in our previous work [23] is used to search the redundant web services with different quality from different WSHs. The WSM is a novel approach implemented as a database that contains various fields such as Service name, Service Cost, Service Quality (frame rate, frame resolution), Host-ID, Host-Status, Service Duration. The sample Web Service Map is shown in Table I. The hosts which provide the requested service are specified with their Host-ID in the corresponding fields. The host status is used to specify whether the host is in busy state (processing a previous client request) or available state (free to accept client requests). The WSM is to be updated periodically to ensure that up-to-date information about the web services and their availability can be obtained by the user and server. The activation and deactivation of web services in WSHs also update WSM to indicate the web service availability.

TABLE I
WEB SERVICE MAP (WSM)

| Service name | Service Cost | Service Quality | | Host-ID | Host-Status | Service Duration (mins) |
		Frame rate(fps)	Frame Resolution (pixels)			
S1	100$	30	720 x 576	WSH-1	Busy	30
S1	120$	15	176X144	WSH-2	Available	50
S2	50$	50	720 x 576	WSH-2	Available	30
S2	60$	30	176x144	WSH-3	Busy	55
S3	75$	15	176X144	WSH-3	Busy	60
S3	90$	50	1024×768	WSH-1	Available	20

D. Media Content Adapter (MCA)

When the requested service cannot be provided with the requested quality, the QoS renegotiation with the user is performed by AMSM. Based on the renegotiation results, the possible media content adaptation is performed by MCA. The MCA modifies the available multimedia content according to the user's display device using three important translation services such as transcoding, scaling and bit-depth reduction as shown in Figure 2. To adapt a high resolution MPEG video stream with 720X576 pixels, 16 bits/pixel, and 30 frames/sec to the low resolution device such as mobile phone, three stages of adaptation is carried out. In the first stage, scaling service reduces the size to 176X144 pixels, the next stage transcoding process decodes the stream and re-encodes it using different codec to get H.264 video, and the last stage is specialized in adapting the color information to 8-bit for the limited capabilities of the mobile device. Such adapted multimedia content is delivered to users and stored in the local multimedia database.
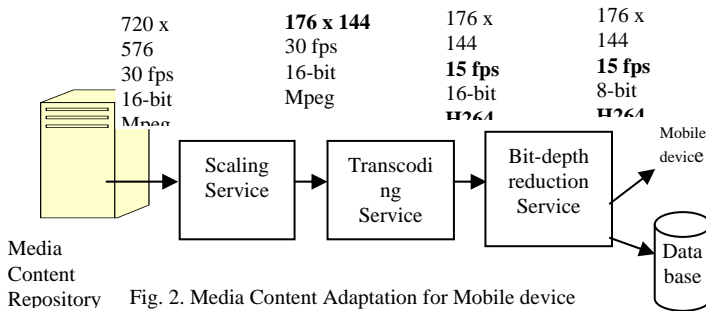
Fig. 2. Media Content Adaptation for Mobile device

E. Adaptive Media Service Manager (AMSM)

The AMSM is designed as a Quality Enforcement Center (QEC) by the enactment of adaptation policies implemented using the run-time components such as MSM, MRA, QR, LB, APR, and APP. The AMSM is the self-adaptive execution engine which monitors, analyses and takes corrective actions to make the system self-adaptive. When changes in quality are detected by MSM during run-time due to server overload, service fault, and SLA violation, the renegotiation with the user has been done before taking the corrective action. If the response time increases above the threshold wait interval 5 seconds, then the AMSM retries the same service for a particular number of times based on the adaptation policy without informing to the user. Such corrective actions are taken by AMSM to give the illusion that no performance degradation is identified by the users.

Functions of AMSM:

- Monitor the composition and run-time performance of multimedia services
- Fault analysis based on the monitored results
- Enforce corrective action by parsing the Adaptation Policy
- Perform QoS renegotiation with the service user for adaptation

The descriptions of its run-time components are given below:

1) Media Service Monitor (MSM)

The Media Service Monitor is the component which is responsible for self-adaptive system that monitors the run-time performance of the media web services. The Parallel Performance Monitoring Service [2] proposed in our previous work is a monitoring web service that can be run along with the media service using multithreading technology which monitors the response time, timeout, percentage of successful completion of media web services and external errors. In addition to these quality metrics, the MSM checks the reliability, availability of media web services, tracks the performance of the media server, monitors the frame rate and measures the business metric called reputation. The eight key QoS metrics and one business metric that are monitored are listed as follows:

a) Response time, calculated using software wrapping technique

b) Reliability, calculated as a ratio of successful invocations over the number of total invocations in given period of time

c) Availability, calculated as the percentage of time that a service is available for a specific interval of time

d) Percentage of successful completion, calculated using polling technique

e) External Errors detection, using raised exceptions

f) Time at which web server response time starts to degrade.

g) Threshold over QoS guarantees are compliance with pre-established SLAs

h) Frame rate

The sample results of the MSM in the cases of increased response time and interruption by user are shown in the figure 3(a) and 3(b) respectively.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<ProcessId>java.lang.ProcessImpl@2acfa2 </ProcessId>
<MediaFile>AVSEQ08.dat</MediaFile>
<WaitInterval>3 Sec </WaitInterval>
<ResponseTime>4 Sec </ResponseTime>
<timeout>expires</timeout>
```

Fig. 3(a). Increased response time

```xml
<?xml version="1.0" encoding="UTF-8"?>
<ProcessId>java.lang.ProcessImpl@1e295f8 </ProcessId>
<MediaFile>DevaUmSachinan.dat</MediaFile>
<WaitInterval>3 Sec </WaitInterval>
<ResponseTime>1 Sec </ResponseTime>
<PlayDuration> 10 Sec</PlayDuration>
<Termination>Interrupted by the user </Termination>
```

Fig. 3(b) Interruption by user

2) Monitored Results Analyzer (MRA)

The Monitored Results Analyzer analyzes the output of the MSM to recognize the type of fault such as timeout, interruption by user, service failure, service unavailable fault, SLA violation, web server overload and network fault. With the type of fault monitored and identified, AMSM performs corrective actions that are listed in Table II based on the policies available in Adaptation Policy Repository. There are different types of policies available, but the policies that are used in this paper based on Event-Condition-Action rules (ECA). The ECA rule normally specifies a triggering event, conditions to be satisfied and actions to be taken.

TABLE II
FAULT TYPE AND CORRECTIVE ACTION TO BE TAKEN

S.No.	Identified Fault type	Corrective Action
1.	Timeout expires (response time>=3s)	Retry
2.	Interruption by user	Next time the same user logs in, display the previous viewed files and allow them to see again
3.	Service Failure Fault	Substitute same type of service from different host
4.	Service Unavailable Fault	SLA generation, if accepted by user, provide same type of service with available quality
5.	SLA violation Fault	QoS renegotiation and adaptation
6.	Web server over load error	Substitute same type of service from different host if available;otherwise retry after some time.
7.	Service failure in between	Restart from the point at which the fault is detected or user interrupted
8.	Network fault	Start the service after some period of time
9.	External fault	Retry

3) QoS Renegotiator (QR)

The QoS provided by the multimedia applications are prone to vary based on network conditions. To provide optimal quality service, the dynamic change in QoS parameters is to be managed. The two important steps such as notification of change and adaptation or renegotiation of QoS parameters are to be done during run-time. The renegotiation is done at run-time when any violation in minimum negotiated value (QoSmin) or changes in the negotiated range (QoSmin, QoSmax) is detected. These violations are monitored by the MSM and the QoS Renegotiator updates the negotiated QoS profiles at run-time and informs to the user. To provide better quality services, the QoS adaptation is done through adaptation policies. For example, Adaptation Policy Repository for response time assertion is checked by AMSM, when the response time is more than the wait interval, to retry the service based on the details available in the policies.

4) Load Balancer (LB)

To balance the load of the media server, redundant media services with different quality are deployed in different Web Service Hosts (WSHs). Adaptive Service Manager checks the load of available WSHs such as number of requests serviced, processor speed, and available memory size using load balancer. The WSH which is servicing less number of requests, using less amount memory is selected to service the current request.

5) Adaptation Policy Repository (APR)

When the self-adapting architecture is required for customization of composition and guaranteeing reliable services, it is advantageous to externalize the descriptions of actions to be taken for individual failure cases from the description of the base process. This separation of concerns for distributed systems is achieved using Policy-based Management [24]. The policies are used for the representation of all types of adaptation activities. The general definition of 'policy' is that it is declarative, high-level description of goals to be achieved and actions to be taken in different situations. The main advantage of policies compared to aspect-oriented programming is that policies are higher-level abstractions and can be specified more easily. In this paper, the Adaptation Policies are developed in a XML format and stored in Adaptation Policy Repository. The sample Adaptation Policy is shown in figure 4 which is an extension of the Web

Services Policy Framework (WS-Policy) [20] which is used to enable specification of policies for self-adapting the architecture when unforeseen faults are detected at run-time. The sample events and its corresponding actions to be taken are specified in the policies as follows:

a) If response time is greater than threshold value; then the RPAssertion is checked for retrying the service based on the information available in the policy as shown in the figure 4.

b) If not successful with retries and external errors are occurred, then the ERAssertion is checked for handing errors by substituting with alternate service with same quality or reduced quality by renegotiation with client during run-time.

c) If the problem continues, then the SKAssertion is checked to skip the service at current point of time and then retry after some time.

```
<wsp:All>
<wsp:AdaptivePolicy Name="tns:RetryAdaptivePolicy" policyType="Retry"
        wsp:Preference="50">
  <wsrp:RPAssertion Name="ResponseTime">
    <wsrp:MaxNumImmediateRetries Value="1"/>
    <wsrp:MaxRetryCycles Value="2"/>
    <wsrp:DelayBetweenRetryCycle Millisec="3000"/>
    <wsrp:MaxNumRetriesPerCycle Value="2"/>
    <wsrp:NotifySenderAfterLastRetryFail Value="1"/>
  </wsrp:RPAssertion>
  <wsep:ErrorAssertion>  <wsep:terminate />   </wsep:ErrorAssertion>
</wsp:AdaptivePolicy>
</wsp:All>
```

Fig. 4. Sample Adaptation Policy

6) Adaptation Policy Parser (APP)

The Adaptation Policy Parser which is implemented as the XML Parser that allows the AMSM to read and understand the Adaptation Policy Repository. When the MSM alerts for an error, the adaptive manger triggers the Adaptation Policy Parser to parse the Adaptation Policy Repository to get the required information about the corrective action to be taken. Its instance is created during run time when it is triggered by adaptive manager.

IV. CASE STUDY- DCRVWS

Policy-based SAMSA supports for customization and guaranty of quality services has been evaluated and demonstrated in various adaptation scenarios using the DCRVWS case study that is implemented as Java based application. The Dynamic Composite Real-time VoD Web Service (DCRVWS) is designed as a business process using BPEL Designer as shown in figure 5. The BPEL process includes the sub processes such as user profile manager, user preference collector, service retriever from WSM, media content adapter and SLA generator.
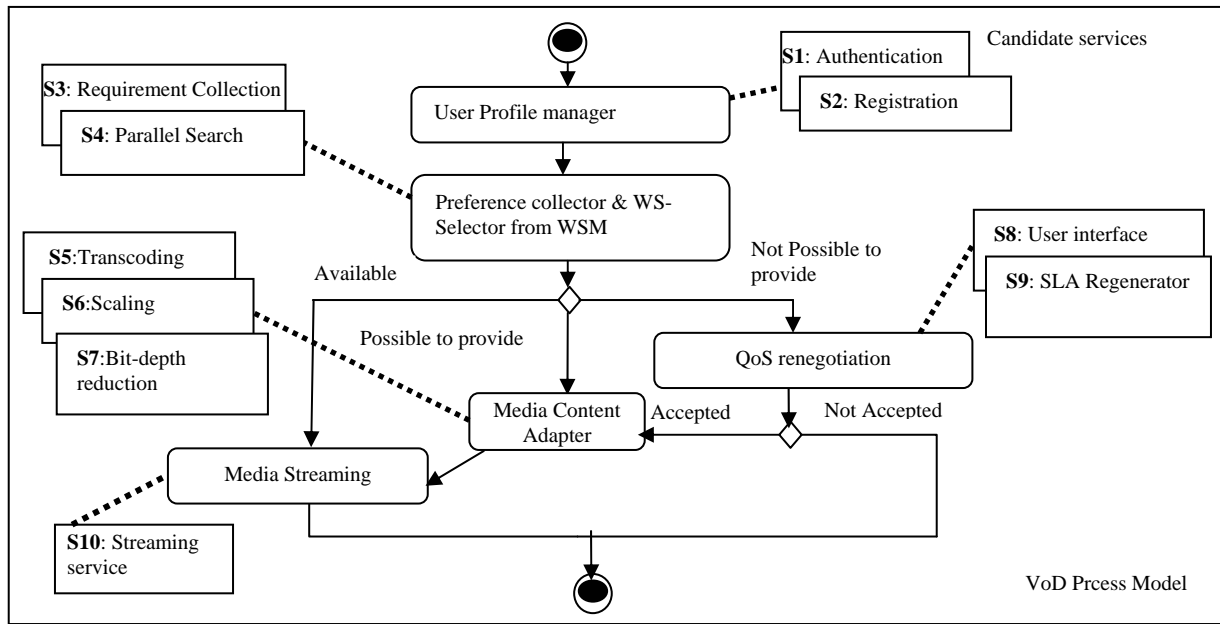
Fig. 5. Dynamic Composite Real-time VoD Web Service (DCRVWS)

According to the user preferences such as media quality, media type, software availability and the type of device used, the composition of services are done dynamically.

Sample DCRVWS:

Case 1: For instance the user requesting for video at first time with parameters such as cost range, frame rate, frame resolution, device type and software used and assuming that the same is available with requested quality. The services such as registration (s2), requirement collection (s3), parallel search (s4) and streaming service (s10) are invoked. DCRVWS will be: s2→s3→s4→s10

Case 2:
If the requested service is not available, service level agreement (SLA) is drafted between the service provider and service user by invoking SLA service S9. Based on the agreement, the available service is invoked. Now the DCRVWS will be: s1→s3→s4→s8→s9→s10

Case 3:
When the service is interrupted in between and the same user logs in next time then the information about the interrupted video service such as name of the service, quality parameters and how much time the video is viewed has been listed. The user is allowed to view the interrupted video from the point at which it was interrupted. This time the DCRVWS will be: s1→s10

Case 4:
If the requested video service with the quality such as 176 x 144 pixels resolution, 15 fps, 8-bit, H.264 is not available and possible to provide with media content adaptation services such as transcoding (s5), scaling (s6) and bit-depth reduction (s7), then MCA component is invoked. The MCA modifies the available video according to the user requirements. Now the DCRVWS will be: s1→s3→s4→s5→s6→s7→s10.

The important run-time component Media Service Monitor is parallely invoked with media web services such as s5, s6, s7 and s10. It monitors the run-time performance of the media services and the monitored results are analyzed and corrective actions are taken by AMSM based on policies.

V. EVALUATION

To evaluate the effectiveness of this proposed architecture, the experiment has been conducted using case study DCRVWS. The web services are developed and deployed in the Java-based environment. ActiveBPEL designer and execution engine have been used to design the composition of web services to service the VoD request. The video web services with different quality are deployed into cluster of PCs that have same configuration. The information about web services is available in Web Service Map that is used as registry. The experiments have been conducted using 200 client machines that are connected to a LAN through 100Mbps Ethernet cards. At a time 200 client requests have been generated and the results of service deliveries are analyzed.

The technical quality metrics such as Reliability (R_e), and Response time (R_t) and the business metric Reputation (R) are used to analyze the effectiveness of the proposed architecture. Reliability $R_e(s)$ of a media service is defined as the probability that a service is continuously and correctly delivered with in the maximum expected time frame indicated by the threshold value. Especially, the reliability of media service includes the two important characteristics such as continuous and correct delivery of service within expected response time. The AMSM monitors these run-time characteristics through MSM and if any deviation occurs due to external errors, then AMSM takes corrective actions such as retry or substitute a faulty service with its equivalent service,. To ensure the service delivery with in the specified response time, the MSM retries the same service. The table III shows the improvement in response time when the self-

adaptation is done through corrective actions specified by policies.

TABLE III
ANALYSIS OF RESPONSE TIME

No. of Simultaneous requests	Average Response Time (ms)	
	Without adaptation	With self-adaptation
25	80	60
50	100	80
75	150	120
100	200	150
125	400	250
150	600	450
175	920	700
200	1200	980

To ensure continuous delivery of multimedia service, the MSM which runs parally with media service, monitors the media service through polling technique. Using this technique, the MSM polls every 2 minutes and checks whether service is continuously delivered or not. When the MSM finds that the service is interrupted in the middle then it is notified to AMSM. The AMSM informs the client about the interruption and allows the user to see the same video service from where it is interrupted. This sample output screen is shown in figure 6(a) and 6(b). In this way the reliability can be achieved using this proposed architecture. The value of the reliability of a media service is computed from the data of past invocations using the expression $R_e(s) = N(s)/K$, where $N(s)$ is the number of times the service 's' has been successfully delivered within the specified expected time frame, and K is the total number of invocations. By aggregating the reliability of individual services, the reliability of DCRVWS is computed. The figure 7 shows that the reliability of the composition and media services that are improved approximately 20% with self-adaptation compared with no adaptation.
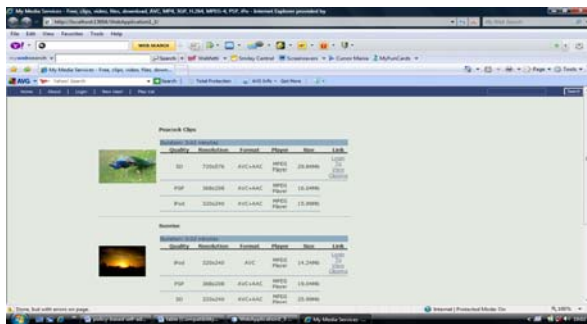
Fig. 6(a): Sample Output Screen – user requirement collection
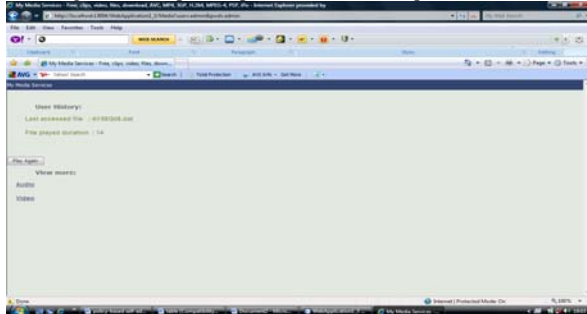
Fig. 6(b). Sample Output Screen – displaying the interrupted video
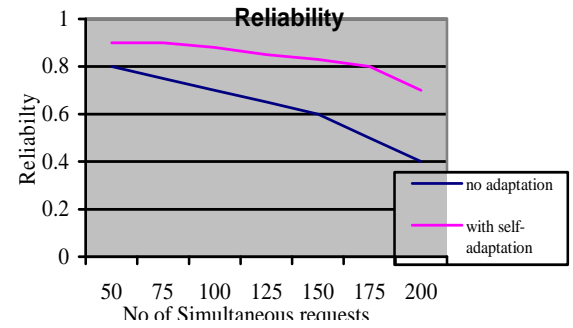
Fig. 7. No. of simultaneous requests vs Reliability

The business metric Reputation R(s) of a multimedia service is a measure of its trustworthiness and it depends on the user's experiences of using the service 's'. Different users may have different opinion on the same service. At the end of the usage of services, the user is given a range [0,5] to rank a service. The numeric value for the reputation of a service is computed by taking the average of rank given by the users as follows:

$$R(s) = \frac{\sum_{i=1}^{n} Ri}{n} \text{---------(1),}$$

Where Ri is the user's ranking on a service's reputation, n is the number of times the service has been graded. The experiment has been conducted to access the VoD service by 100 different client machines and the grade range given by the users has been recorded and reputation of a VoD service has been computed using the equation(1). With SAMSA, the reputation for a VoD service is improved 20% compared to without any adaptation

VI. CONCLUSION

The Self-Adaptive Media Service Architecture (SAMSA) with corrective adaptation in customization of media web service composition and guarantee quality service provisioning are important techniques towards the creation of agile media business processes. This proposed SAMSA is an extension of SOA and is used to continually adapt to the run-time changes to fulfilling the functional and QoS requirements. By sense-analyze-act method, the service provider monitors the performance of the media web services and analyzes the monitored results for taking corrective action. The case study DCRVWS has been implemented based on this architecture and the effectiveness of the architecture is realized through response time, reliability and user satisfaction analysis.

REFERENCES

[1] McGovern J, Sims O, Jain. A, Little M, ."Enterprise Service Oriented Architectures Concepts, Challenges, Recommendations", *Springer 2006*.

[2] G. Maria kalavathy, P. Seethalakshmi, "Parallel Performance Monitoring service for Dynamically Composed Media Web Services", *Journal of Computer Science,* Vol 5, No. 7, pp. 487 – 492, July 2009.

[3] Verma.K, Sheth.A.P, "Autonomic Web Processes", *in Proceedings of the Third International Conference Service Oriented Computing (ICSOC'05)*, Amsterdam, The Netherlands, LNCS, Vol. 3826, Springer, pp. 1-11, 2005.

[4] The Web Services Interoperability Organization (WS-I). 2003. Supply Chain Management Sample Application Architecture, http://www.ws-i.org/ SampleApplications /SupplyChainManagement/2003-12/SCMArchitecture1.01.pdf

[5] Liu Y, Ngu, A.H.H, Zeng L, "QoS Composition and Policing in Dynamic Web Service Selection," *in WWW, pp. pp.* 66-73, 2004.

[6] Zeng. L, Benatallah. B, Dumas. M, Kalagnanam. J And Sheng. Q.Z, "Quality driven web services composition", *in WWW,* pp. 411-421, 2003.

[7] MSU Video Quality Measurement Tool, Available: http://compression.ru/ video / quality_measure / video _measurement_tool_en.html.

[8] Liguo Yu, " Applying Software wrapping on Performance Monitoring of Web Services", *Journal of Computer Science*, 6, pp. 1-6, 2007.

[9] Khaled Mahbub, George Spanoudakis. "Run-time Monitoring of Requirements for Systems Composed of Web-Services: Initial Implementation and Evaluation Experience", in *Proceeding of the IEEE International Conference on Web Service (ICWS'05) ,pp.* 257-265.

[10] William N Robinson, "Monitoring Web Service Requirements", *In Proceedings of the 11th IEEE International Requirements Engineering Conference,* pp. 65-74, 2003.

[11] Arne Koschel, Irina Astrova, "Event Monitoring Web Services for Heterogeneous Information Systems", in *Proceedings of World Academy of Science, Engineering and technology*. Vol 33, 2008, 50-52.

[12] Onyeka Ezenwoye, S Masoud Sadjadi, "A Proxy-Based Approach to Enhancing the Autonomic Behavior in Composite Services"*, Journal of Networks*, Vol. 3, No.5, pp. 42-53, May, 2008.

[13] Onyeka Ezenwoye, S. Masoud Sadjadi, S.M, "Enabling Robustness in Existing BPEL Processes", i*n Proceedings of the 8th International Conference on Enterprise Information Systems (ICEIS'06),* Paphos, Cyprus.

[14] Onyeka Ezenwoye, S Masoud Sadjadi, "Robust-BPEL: Transparent Autonomization in Aggregate Web Services Using Dynamic Proxies", *Technical Report: FIU-SCIS-2006-06-01.*

[15] Charfi A., Mezini M, "An Apect-Based Process Container for BPEL", in *Proceedings of the First Workshop on Aspect-Oriented Middleware Development (AOMD 2005),* Grenoble, France, ACM.

[16] Baresi L, Guninea S, Plebani P, "WS-Policy for Service Monitoring", in *Proceedings of the 6th International Workshop on Technologies for E-Services (TES 2005),* Trondheim, Norway, Lecture notes in Computer Science (LNCS), Vol.3811. Springer 2005, pp. 72-83.

[17] Birman K, Van Renesse R, Vogels W, "Adding High Availability and Autonomic behavior to Web Services", in *Proceedings of the 26th International Conference on Software Engineering (ICSE'04),* Edinburg, Scotland, UK, pp. 17-26.

[18] Hamadi. R, Benatallah B, "Recovery Nets: Towards Self-Adaptive Workflow Systems", *In Proceedings of the 5th International Conference on Web Information Systems Engineering (WISE '04),* LNCS 3306, pp. 439-453, Springer Verlag, Brisbane, Australia.

[19] Wile. D.S, Egyed A, "An Externalized Infrastructure for Self-Healing Systems", i*n Proceedings of the 4th Working IEEE/IFIP Conference on Software Architecture (WICSA '04),* Oslo, Norway, pp. 285-290, 2004.

[20] IBM et al, "Web Services Policy Framework (WS-Policy), September 2004, http://www.106.ibm.com/developerworks/library/specification/ s-polfram.html.

[21] K. Khitrin et al., Physical Review Lett. 89, 277902 (2002).

[22] Rangeet Bhattacharyya et al, " Implementation of parallel search algorithms using spatial encoding by nuclear magnetic resonance", *Physical Review A* 71, 052313, 2005.

[23] G. Maria kalavathy, P. Seethalakshmi, " Enhancing the Availability of Composite Real-time Multimedia Web Service", i*n Proceedings of the IEEE Workshop on QPMHPC with HPCC'08*, Dalian, China, 1001-1006.

[24] Sloman. M, "Policy-Driven management for Distributed system", *Journal of Network and Systems Management*, Vol.2, No. 4, Kluwer, pp. 333-360, 1994.

Marker-less 3D Human Body Modeling using Thinning algorithm in Monocular Video

[*]K. Srinivasan
Department of EIE
Sri Ramakrishna Engineering College
Coimbatore, India
srineekvasan@gmail.com
* Corresponding author

K.Porkumaran
Department of EEE
Dr.N.G.P Institute of Technology
Coimbatore, India
porkumaran@gmail.com

G.Sainarayanan
Head, R&D
ICT Academy of Tamilnadu
Chennai, India
Sai.jgk@gmail.com

Abstract— **Automatic marker-less 3D human body modeling for the motion analysis in security systems has been an active research field in computer vision. This research work attempts to develop an approach for 3D human body modeling using thinning algorithm in monocular indoor video sequences for the activity analysis. Here, the thinning algorithm has been used to extract the skeleton of the human body for the pre-defined poses. This approach includes 13 feature points such as Head, Neck, Left shoulder, Right shoulder, Left hand elbow, Right hand elbow, Abdomen, Left hand, Right hand, Left knee, Right knee, Left leg and Right leg in the upper body as well as in the lower body. Here, eleven activities have been analyzed for different videos and persons who are wearing half sleeve and full sleeve shirts. We evaluate the time utilization and efficiency of our proposed algorithm. Experimental results validate both the likelihood and the effectiveness of the proposed method for the analysis of human activities.**

Keywords- Video surveillance, Background subtraction, Human body modeling, Thinning algorithm, Activity analysis.

I. INTRODUCTION

In recent years, human tracking, modeling and activity recognition from videos [1]-[5] has gained much importance in human-computer interaction due to its applications in surveillance areas such as security systems, banks, railways, airports, supermarkets, homes, and departmental stores. The passive surveillance system needs more cameras to monitor the areas by a single operator and it is inefficient for tracking and motion analysis of the people for better security. The automated video surveillance system uses single camera with single operator for the motion analysis and provides better results. Marker based human tracking and modeling is a simple way of approach but it is not possible to reconstruct all the human poses in practical situations. This approach needs markers at every time of surveillance persons. So, the marker-less motion tracking and modeling have been very important for the motion analysis. In the human body modeling, there are two kinds of representation of modeling are available such as 2D modeling and 3D modeling. Among the two types of human body modeling, 2D modeling is simple approach which can be used to model the complex nature of human body whereas 3D modeling is much more complex to track the persons in video data. In this paper, 3D human body modeling

based activity analysis has been implemented with the help of thinning algorithm. The recovery of 3D human body poses is a very important in many video processing applications. A 3D human body model is an interconnection of all body elements in three dimensional views. Onishi K.et.al [6] describe a 3D human body posture estimation using Histograms of Oriented Gradient(HOG) feature vectors that can express the shape of the object in the input image obtained from the monocular camera. A model based approach for estimating 3D human body poses in static images have been implemented by Mun Wai Lee, and Isaac Cohen [7]. They develop a Data-Driven based approach on Markov Chain Monte Carlo (DD-MCMC), where component detection results generate state proposals for 3D pose estimation.

Thinning is one of the important morphological operations that can be used to remove the selected foreground pixels from the images. Usually, the thinning operation has been applied to binary images. In the previous work, the thinning algorithm is mostly attempted for several image processing applications like pattern recognition and character recognition [8]-[11]. Now we apply this thinning algorithm to model the human body in 3D view and it can be used to find the motion analysis of human without using any markers on the body.

This paper is organized as follows: Section 1 gives the brief introduction about the problem. Section 2 deals the proposed work of activity analysis using 3D modeling. The frame conversion algorithm and the background subtraction algorithm are explained in section 3 and section 4. Section 5 illustrates the morphological operation and the thinning algorithm is described in section 6. Section 7 presents the human body feature points identification. Section 8 includes the results and analysis of our proposed work. The conclusion and future work has been discussed in section 9. The acknowledgements and references are included in the last part of the paper.

II. PROPOSED WORK

Human body modeling has been used in the analysis of human activities in the indoor as well as in the outdoor surveillance environment. Model based motion analysis involves 2D and 3D human models representation [12]-[13]. The features that are extracted from the human body are useful to model the surveillance persons and it has been applied to

recover the human body poses [14] and finding their activities. In the proposed work as in Figure 1, first the video sequence is acquired by the video camera from the indoor environment and it is converted into frames for further processing. Due to illumination changes, camera noise and lighting conditions, there may be a chance of adding noise in the video data. These unwanted details have to be removed to get the enhanced video frame. The pre-processing stage helps to enhance the video frames. In all the processing here, the human body is our desired region of interest. The next aim is to obtain the human body from the video frame by eliminating the background scene. So, the background is subtracted with the help of the frame differencing algorithm. Then, the video frames are applied to morphological operation to remove details smaller than a certain reference shape. After the morphological operation, the thinning algorithm has been used to find skeleton of the human body. In this work, 13 features have been considered for a full body modeling and these features are Head, Neck, Left shoulder, Right shoulder, Left hand elbow, Right hand elbow, Abdomen, Left hand, Right hand, Left knee, Right knee, Left leg and Right leg as in Figure 2. Initially, the five terminating points such as head, left hand, left leg, right leg, and right hand are determined. Then, the intersecting, shoulder, elbow, and knee points are obtained using image processing techniques. Finally, the 3D modeling has been achieved for the activity analysis of human in video data.

Figure 2.   A human body model with thirteen feature points

III.   FRAME CONVERSION ALGORITHM

In the first stage, the Video sequence is captured by the high resolution Nikon COOLPIX Digital Video Camera having 8.0 million effective pixels and 1/2.5-in.CCD image sensor which produces NTSC and PAL video output. And it has a focal length of 6.3-18.9mm (equivalent with 35mm [135] format picture angle: 38-114mm). The video sequence is being taken at a rate of 30 frames/ second from the indoor surveillance environment for finding the human behaviour. After that, the video sequence has been converted into individual frames with the help of the algorithm given below.

VIDEO TO FRAME CONVERSION ALGORITHM

Step0: Acquisition of video sequence from the Video camera
        to MATLAB environment.
Step1: Read the video file using 'aviread' function and
        store it in a variable name.
Step2: Assign the required frame as 'jpg'.
Step3: Determine the size of video file and number it.
Step4: Then,
        for i=1: fnum,
                strtemp=strcat(int2str(i),'.',pickind);
                imwrite (mov(i).cdata(:,:,:),strtemp);
        end

Figure 1.   Proposed model of 3D modeling for activity analysis

IV.   BACKGROUND SUBTRACTION ALGORITHM

In the proposed work, the background subtraction technique plays an important role for subtracting foreground images from the background image and it is described in Figure 3. The frame differencing algorithm [15] has been

proposed here to highlight the desired foreground scene and it is given below.

FRAME DIFFERENCING ALGORITHM

Step0: Read the Video data.

Step1: Convert it into video frames.

Step2: Set the background image.

Step3: Separate R, G, B components individually for easy computation.

  bc_r = bg(:,:,1);bc_g = bg(:,:,2); bc_b = bg(:,:,3);

Step4: Read the current frame from the video sequence.

Step5: Separate R, G, B components individually for the computation.

  cc_r = fr(:,:,1);cc_g = fr(:,:,2); cc_b = fr(:,:,3);

Step6: Subtract the R, G, B components of the current frame from the R, G, B components of background frame.

Step7: Check the threshold values of colour components.
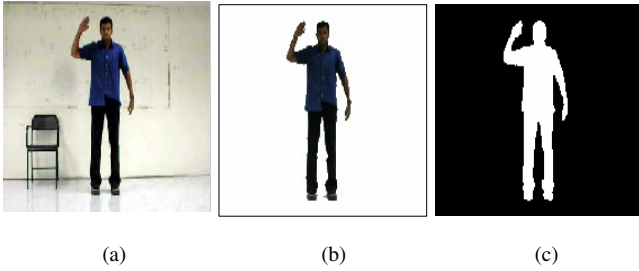
Figure 3.   Background subtraction using frame differencing algorithm
(a) Input video frame, (b) Background subtracted image, and (c) Silhouttee of human body

V.   MORPHOLOGICAL OPERATION

Next, the proposed algorithm follows the morphological operations which help to enhance the video frame for further processes. The morphological operations include dilation and erosion [16]. Finally, the noise has been removed using median filtering. Dilation adds pixels to the boundaries of the objects in an image. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element. If $F(j,k)$, for $1 \leq j,k \leq N$ is a binary valued image and $H(j,k)$, for $1 \leq j,k \leq L$, where L is an odd integer, is a binary valued array called a structuring element, the dilation is expressed as in equation(1).

$$G(j,k)=F(j,k) \oplus H(j,k) \qquad (1)$$

Erosion removes pixels on object boundaries. To erode an image, *imerode* function is used for our applications. The dilation is expressed as in equation (2) where $H(j,k)$ is an odd size Lx L structuring element.

$$G(j,k)=F(j,k) \otimes H(j,k) \qquad (2)$$

At the end of this stage, the median filtering has been used to reduce the salt and pepper noise present in the frame. It is similar to using an averaging filter, in that each output pixel is set to an average of the pixel values in the neighborhood of the corresponding input pixel.

VI.   THINNING ALGORITHM

In this paper, thinning operation can be used to find skeleton of the entire human body. The thinning operation is performed by transforming the origin of the structuring element to each pixel in the image. Then it is compared with the corresponding image pixels. When the background and foreground pixels of the structuring element and an image are matched, the origin of the structuring element is considered as background. Otherwise it is left unchanged. Here, the structuring element determines the use of the thinning operation. The thinning operation is achieved by the hit-and-miss transform. The thinning of an image A by a structuring element B is given by equation (3).

$$\text{thin}(A,B)=A\text{-hit and miss}(A-B) \qquad (3)$$

Mostly the thinning operation has been used for skeletonization to produce a connected skeleton in the human body. Figure.4 shows the structuring elements for skeletonization by morphological thinning. At each iteration, the image is first thinned by the left hand structuring element, and then by the right hand one, and then with the remaining six 90° rotations of the two elements.

Figure 4.   Examples of structuring element for thinning operation

The process is repeated in cyclic fashion until none of the thinnings produce any further change. Normally, the origin of the structuring element is at the center. The steps of thinning algorithm include,

Step0: Partitioning the video frame into two distinct subfields in a checkerboard pattern.

Step1: Delete the pixel *p* from the first subfield if and only if the conditions (4), (5), and (6) are satisfied.

$$X_H(p)=1 \qquad (4)$$

$$X_H(p)=\sum_{i=1}^{4} b_i$$

where

$$b_i = \begin{cases} 1 \text{ if } X2i\text{-}1 = 0 \text{ and } (x2i = 1 \text{ or } x2i+1=1) \\ 0 \text{ otherwise} \end{cases}$$

$x1, x2,….x8$ are the values of the eight neighbors of *p*, starting with the east neighbor and numbered in counter-clockwise order.

$$2 \leq \min\left\{n_1(p),n_2(p)\right\} \leq 3 \qquad (5)$$

$$n_1(p) = \sum_{i=1}^{4} X_{2k-1} \vee X_{2k}$$

where

$$n2(p) = \sum_{i=1}^{4} X_{2k} \vee X2k+1$$

$$(X2 \vee X3 \vee \overline{X\,8}) \wedge X1 = 0 \qquad (6)$$

Step2: Then, delete the pixel *p* from the second subfield if and only if the conditions (4), (5), and (7) are satisfied.

$$(X6 \vee X7 \vee \bar{X}4) \wedge X5 = 0 \qquad (7)$$

The combination of step1 and step2 produce the one iteration of the thinning algorithm. Here, an infinite number of iterations ($n=\infty$) have been specified to get the thinned image. Figure.5 shows the thirteen points on thinned image for different poses.
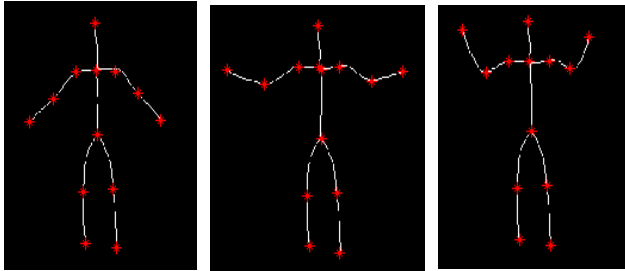
Figure 5.   Results of thinned image for a human body with 13 points

VII.   HUMAN BODY FEATURE POINTS IDENTIFICATION

In order to model the human body, thirteen feature points have been considered in the upper body as well as the lower body. The feature points include the Terminating points (5Nos), Intersecting points (2Nos), Shoulder points (2Nos), Elbow joints (2Nos), and Knee joints(2Nos).Using terminating points, the ends of features such as head, hands and legs have been determined. The following are the steps involved in determining the terminating points.

STEPS TO FIND TERMINATING POINTS

Step0: Input the thinned image.
Step1: Initialize the relative vectors to the side borders from the current pixel.
Step2: Select the current coordinate to be tested.
Step3: Determine the coordinates of the pixels around this pixel.
Step4: If this pixel is an island, then it is an edge to the island of 1 pixel. Save it.
Step5: Default assumption: pixel is an edge unless otherwise stated.
Step6: Test all the pixels around this current pixel.
Step7: For each surrounding pixel, test if there is a corresponding pixel on the other side.
Step8: If any pixels that are on the opposite side of the

current pixel, then this pixel is not a terminating pixel.
Step9: If the current pixel does not satisfy the above condition, then it is an edge.

Once the terminating points are determined, then the two intersecting points have been calculated which joints hands and legs. Then, the two shoulder points are determined. The left shoulder co-ordinate is plotted at the pixel where the iteration encounters a white pixel. Similarly the right shoulder co-ordinate is plotted using the same technique. Figure.6 shows a graphical representation to determine Shoulder, Elbow and Knee of the human body.
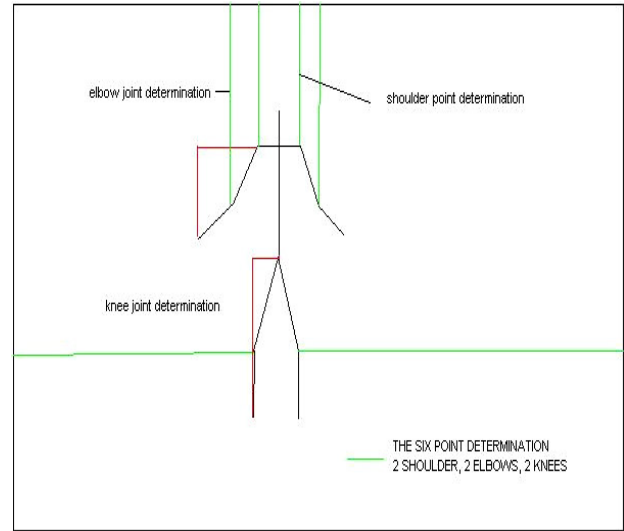
Figure 6.   Graphical representation to find Shoulder, Elbow and Knee

The elbow point is approximately halfway between the shoulder and the terminating points of the two hands. The problem arises when the hand is bent. In order to get the accurate elbow joint, a right angle triangle has been constructed as in Figure 7(a).The (x1, y2) point of the right angled triangle is determined by obtaining the x-axis of the terminating point-1 (x1) and the y-axis of the shoulder point (y2). The distance between the point (x1, y1) and (x2, y2) is calculated by using the equation (8).

$$\text{Distance between points (D)} = \sqrt{(x_1 \text{-} x_2)^2 + (y_1 \text{-} y_2)^2} \qquad (8)$$

$$\text{Elbow Joint (EJ)} = \frac{\sqrt{(x_1 \text{-} x_2)^2 + (y_1 \text{-} y_2)^2}}{2} \qquad (9)$$

Using the available distance as the x-axis reference, a *for loop* is iterated from the first point of the same x-axis. The point at which the iteration encounters a white pixel is plotted as the elbow joint. Similarly, the other elbow joint is also determined using the same technique. The process of determining the knee joints is similar to the technique adopted to determine elbows. Figure 7(b) shows the graphical way to

determine the knee joint. But, in this case the loop is iterated with a constant y-axis and a varying x-axis. The elbow joint has been identified using the equation (9). After the determination of thirteen points, it has been displayed as in Figure.5.

(a)   (b)
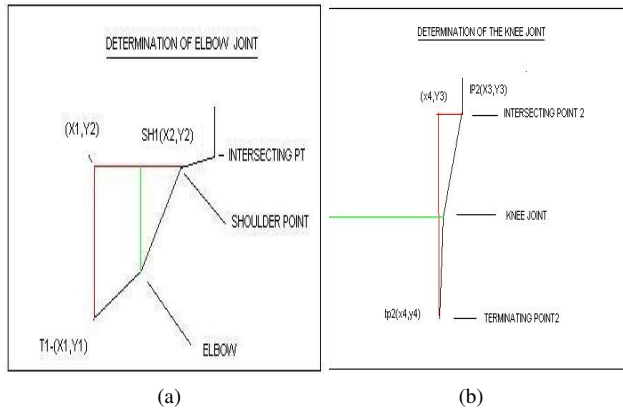
Figure 7.   Graphical representations to find Elbow joint and Knee joint
(a) Determination of Elbow joint (b) Determination of Knee joint

VIII.   RESULTS AND ANALYSIS

In this section, the experimental results of the proposed work are shown and the algorithm has been developed using MATLAB 7.6(2008a) on Intel dual core processor, 2 GB RAM and Windows XP SP2. For implementing this 3D human body model, more than 60 videos are considered. Figure.8 shows the MATLAB results of human body modeling in a 3D view for a single person with different views.
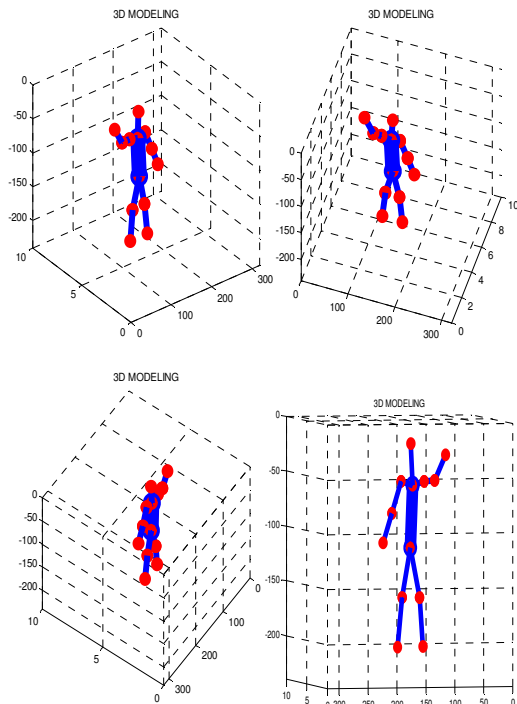
Figure 8.   Results of 3D Modeling of human pose in a different views

This algorithm is implemented for a single person in indoor surveillance video with straight poses for eleven activities such as Standing, Right hand rise, Left hand rise, Both hands rise, Right hand up, Left hand up, Both hands up, Left leg rise, Right salute, Left salute, and Crouching as in Figure 9.

A

B

C

D

E

F

G

H

I

J

K
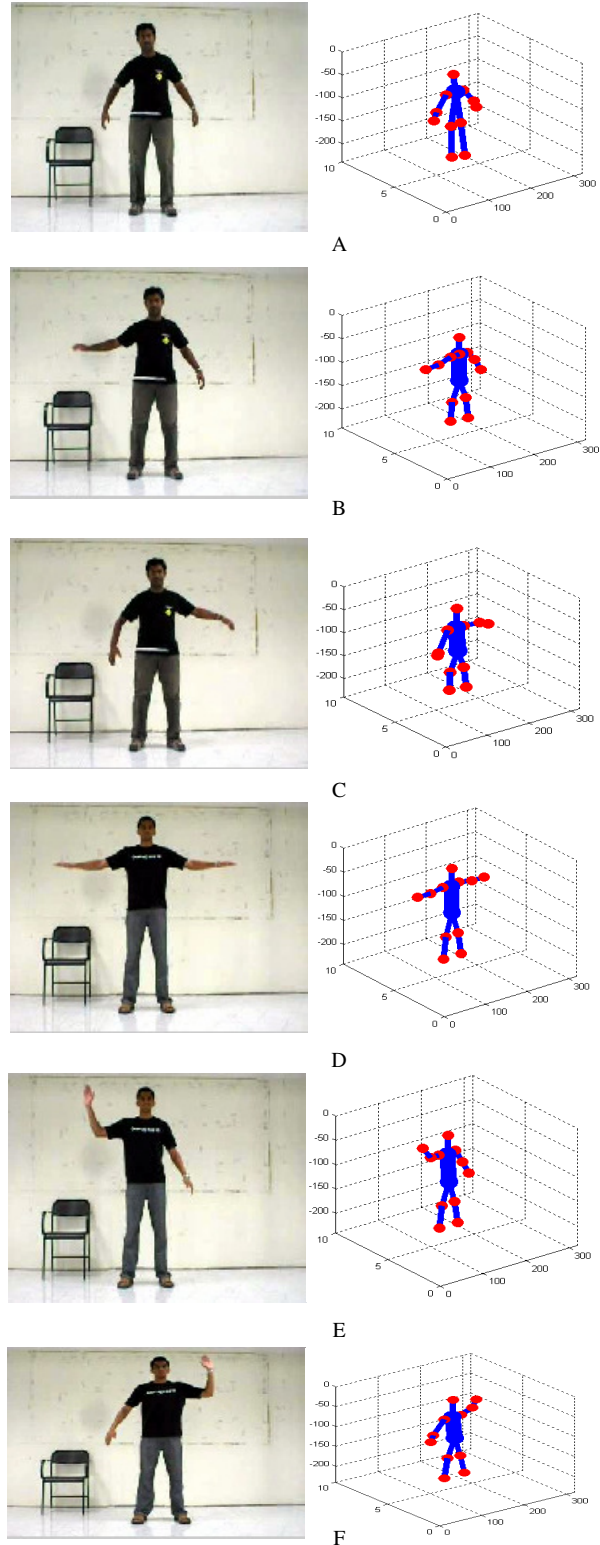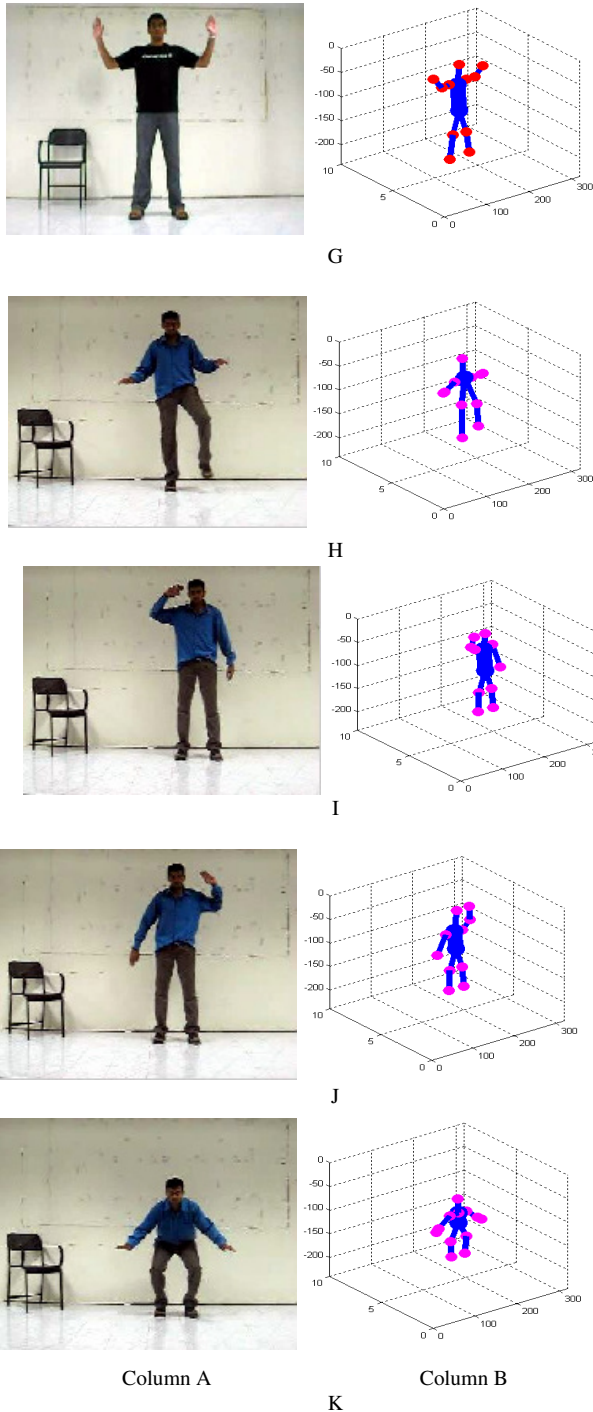
Column A   Column B

Figure 9.   Experimental results of Activity analysis using 3D modeling for different persons.
(Column A) Original video frame   (Column B) 3D modeling
A. Standing, B.Right hand rise, C.Left hand rise, D.Both hand rise,E. Right hand up,F.Left hand up ,G.Both hands up, H. Left leg rise,I. Right salute, J. Left salute, and K. Crouching

To post process the frames for the identification of human activities, silhouette matching technique is used. For this, the silhouettes of eleven activities are stored in the data base. Then, the thirteen feature points of current video frame are identified and compared with the silhouette of the human body

model. If that thirteen points are matched and inside of silhouette, then the corresponding activity is identified. From the response shown in Figure.10, the time taken to compute our algorithm with the steps of 10 frames for a video is observed. For a first frame in the video sequence, it takes approximately 2.6 seconds as high as compared to consecutive frames due to the computation of initial processing like frame conversion, background subtraction and preprocessing. It was noticed that the proposed algorithm has taken 1.6 seconds as an average for 3D models.
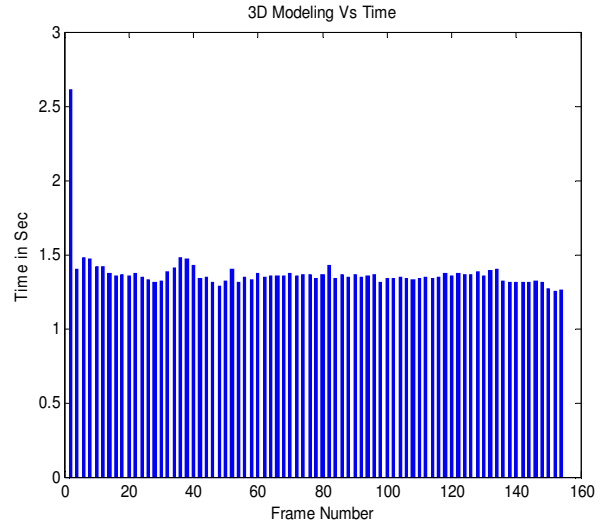
Figure 10. Response of time utilization for an indoor video

We have experienced in the proposed models with eleven activities as in Table I in the indoor monocular videos. Here, we have considered three videos for calculating the algorithm speed of our proposed models. For Video1, it takes an average of 1.62 seconds, and 1.68, 1.78 for the video2 and video3 respectively. The efficiency of our models has been found based on the True positives (TP) and False positives (FP). True Positives indicate the number of frames in which the output is correct in a video sequence. False Positive is the number of frames for which the output is incorrect. Table II shows the efficiency of our proposed modeling for different videos.

TABLE I.   TIME CALCULATION OF ELEVEN ACTIVITIES

S.No	Activity Name	Video 1 (Sec)	Video 2 (Sec)	Video 3 (Sec)
1	Standing	1.81	2.01	2.25
2	Right hand rise	1.74	1.89	2.05
3	Left hand rise	1.59	1.78	1.64
4	Both hand rise	1.76	1.64	2.00
5	Right hand up	1.58	1.82	1.66
6	Left hand up	1.57	1.52	1.56
7	Both hands up	1.56	1.59	2.00
8	Left leg rise	1.54	1.54	1.50
9	Right salute	1.63	1.65	1.69
10	Left salute	1.58	1.71	1.58
11	Crouching	1.50	1.40	1.66
	Average	1.62	1.68	1.78

TABLE II.　　EFFICIENCY OF OUR PROPOSED ALGORITHM

Input	TP	FP	TP+FP	TP/(TP+FP)	Efficiency (%)
Video 1	914	92	1006	0.9085	90.85
Video 2	1100	47	1147	0.9590	95.90
Video 3	1286	96	1382	0.9305	93.05
Video 4	798	66	864	0.9236	92.36
Video 5	1349	153	1502	0.8981	89.81
Video 6	1114	143	1257	0.8862	88.62
Video 7	1171	115	1286	0.9105	91.05

IX. CONCLUSION AND FUTURE WORK

We have implemented an approach for Human 3D modeling for the motion analysis in video security applications. The proposed algorithm works on straight poses acquired by single static camera without using markers on the human body. Here, eleven activities of 3D models have been discussed based on the thinning algorithm and these activities are used to describe almost all human activities in the indoor environment. We have considered 13 feature points for the upper body modeling as well as for lower body modeling. In this paper, time expenditure and efficiency of pre-defined 3D models have been presented. In the future work, this work can be extended to develop an algorithm for multiple persons tracking and modeling. Here, the occlusion problem of human body segments is not considered. This problem will also be considered with outdoor surveillance videos with side poses.

ACKNOWLEDGMENT

REFERENCES

[1] N.Jin, F. Mokhtarian, "Human motion recognition based on statistical shape analysis," Proceedings of AVSS, pp. 4-9, 2005.

[2] Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang, "Human activity detection and recognition for video surveillance," Proceedings of ICME, Vol. 1, pp. 719-722, 2004.

[3] H.Su, F. Huang, "Human gait recognition based on motion analysis," Proceedings of MLC, pp. 4464-4468, 2005.

[4] Tao Zhao, Ram Nevatia and Bo Wu, "Segmentation and Tracking of multiple humans in crowded environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, No. 7, pp.1198-1211, July 2008.

[5] Mun Wai Lee, and Ramakant Nevatia, "Human Pose Tracking in monocular sequence using multilevel structured models," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 1, pp.27-38, 2009.

[6] K.Onishi, T.Takiguchi, and Y.Ariki, "3D Human posture estimation using the HOG features from monocular image," Proc. of 18th IEEE Int. conference on Pattern Recognition, Tampa, FL, pp.1-4, 2008.

[7] Mun Wai Lee, and Isaac Cohen, "A Model based approach for estimating human 3D poses in static Images," Trans. on Pattern Analysis and Machine Intelligence, Vol.28, No.6, pp.905-916, June 2006.

[8] S.Veni, K.A.Narayanankutty, and M.Kiran Kumar, "Design of Architecture for Skeletonization on hexagonal sampled image grid," ICGST-GVIP Journal, Vol.9, Issue (I), pp.25-34, February 2009.

[9] L Huang, G Wan, and C Liu "An improved parallel Thinning algorithm," Proc. of the Seventh International conference on document analysis and recognition ,Vol.2, pp. 780-783, 2003.

[10] V.Vijaya Kumar, A.Srikrishna, Sadiq Ali Shaik, and S. Trinath "A new Skeletonization method based on connected component approach" IJCSNS Int.J. of Computer Science and Network Security, Vol.8, No.2, pp.133-137, February 2008.

[11] S. Schaefer and C. Yuksel, "Example-Based Skeleton Extraction", Proc. of Eurographics Symposium on Geometry Processing, pp. 1–10, 2007.

[12] R.Horaud, M.Niskanen, G. Dewaele,and E.Boyer, "Human motion tracking by registering an articulated surface to 3D points and normals," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, pp. 158-163, 2009.

[13] Jianhui Zhao, Ling Li and Kwoh Chee Keong, "Motion recovery based on feature extraction from 2D Images," Computer Vision and Graphics, pp. 1075–1081,Springer, Netherlands. , 2006.

[14] Jingyu Yan, M.Pollefeys, "A Factorization based approach for articulated non-rigid shape, motion and Kinematic chain recovery from video," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, Issue 5, pp. 865-877, 2008.

[15] K.Srinivasan, K.Porkumaran, G.Sainarayanan, "Improved background subtraction techniques for security in video applications," Proceedings of 3rd IEEE International Conference on Anti-counterfeiting, Security, and Identification in Communication, pp.114-117, 20-22 Aug. 2009.

[16] William K. Pratt, "Digital Image Processing", Jhon Wiley & Sons, Inc., Third edition, 2002.

AUTHORS PROFILE

K.Srinivasan received his BE degree in Electronics and Communication Engineering from VLB Janakiammal College of Engineering and Technology, Coimbatore and ME in Process Control and Instrumentation Engineering, from Annamalai University, India in 1996 and 2004 respectively. He is currently working as an Assistant Professor at Sri Ramakrishna Engineering College, Coimbatore, India. His research interest includes Image/Video Processing, Digital Signal Processing and Neural Networks and Fuzzy systems.

K.Porkumaran is a Vice-Principal in Dr. N.G.P. Institute of Technology, Anna University, Coimbatore, India. He received his Master's and PhD from PSG College of Technology, India. He was awarded as a Foremost Engineer of the World and Outstanding Scientist of the 21st Century by the International Biographical Centre of Cambridge, England in 2007 and 2008 respectively. He has published more than 70 research papers in National and International Journals of high repute. His research areas of interest include Image and Video processing, Modelling and Simulation, Neural Networks and Fuzzy systems and Bio Signal Processing.

G.Sainarayanan received his Engineering degree from Annamalai University, and ME degree from PSG College of Technology, India in 1998 and 2000 respectively and PhD degree from School of Engineering and Information Technology, University Malaysia Sabah, Malaysia in 2002. He is currently working as a Head of R&D, ICT Academy of Tamilnadu, Chennai, India. He is an author of many papers in reputed National and International journals and he has received funds from many funding agencies. His research areas include Image/ video processing, Video Surveillance Systems, Control Systems, Neural Network & Fuzzy Logic, and Instrumentation.

Cryptanalysis on Two Multi-Server Password Based Authentication Protocols

Jue-Sam Chou[*]
Dept. of Information Management
Nanhua University, Taiwan
jschou@mail.nhu.edu.tw
[*]: corresponding author

Chun-Hui Huang
Dept. of Information Management
Nanhua University, Taiwan
g6451519@mail.nhu.edu.tw

Yalin Chen
Institute of Information Systems and
Applications, NTHU, Tawain
d949702@oz.nthu.edu.tw

Abstract¡ In 2004 and 2005, Tsaur et al. proposed two smart card based password authentication protocols for multi-server environments. They claimed that their protocols are safe and can withstand various kinds of attacks. However, after analyses, we found both of them have some security loopholes. In this article, we will demonstrate the security loopholes of the two protocols.

Keywords- multi-server; remote password authenticationl; smart card; key agreement; Lagrange interpolating polynomial

I. INTRODUCTION

In a traditional identity authentication mechanism, a user must use his identity ID and password PW to register at the remote server and the server needs to employ a verification table to record the ID and PW. However, this approach might make the system suffer from the stolen verifier attack. To address this problem, some researchers suggested the authentication system adopt a non-verification-table approach. In 1990, Hwang et al. [4] first proposed a smart card based authentication protocol by using such a non-verification-table way. Thereafter, many smart-card non-verification-table based authentication schemes [1, 2, 3, 5, 6, 7, 10-20] were proposed. In 2004 and 2005, Tsaur et al. proposed two such authentication schemes [8, 9] for multi-server environments. They claimed that their schemes are secure and can withstand various attacks. However, after analyses, we found that both of them have some security loopholes. In this article, we will demonstrate the security flaws found in their protocols.

II. REVIEW AND ATTACK ON TSAUR ET AL.¡S FIRST PROTOCOL

A. Review

Tsaur et al.¡s first protocol [8] consists of next four stages.

a) The System Setup Stage: CA defines an one-way hash function $h(X, Y)$; he selects two large prime numbers p_1, p_2, and computes $N = p_1 ¡ p_2$; he randomly chooses the encryption key e satisfying $\gcd(e, \varphi(N)) = 1$, where $\varphi(N) = (p_1 ¡ 1) ¡ (p_2 ¡ 1)$, and computes his corresponding private key as $d = e^{-1} \mod \varphi(N)$. For each server S_j, CA selects a random S_SK_j as the server¡s private key and computes $S\_ID_j = g^{S\_SK_j} \pmod N$ as his oublic identity, where $j = 1,2, ..., m$.

b) The User Registration Stage: When a new user U_i wants to register at m servers, S_1, S_2, ¡, and S_m (in a multi-server system), he and CA together perform the registration process through a secure channel described as follows:

- U_i chooses his identity U_ID_i and password U_PW_i, and transmits them to CA.

- CA randomly chooses a number r_{ui}, and computes two secret keys as

  $U\_R_i = g^{U\_PW_i * r_{ui}} \pmod N$ and

  $U\_S_i = g^{r_{ui} * d} \pmod N$.

- CA assumes that U_i wants to obtain the services of r servers, $S_1, S_2, ¡, S_r$, for $1 \le r < m$. The service periods provided by these servers are E_T_{i1}, E_T_{i2}, ¡, and E_T_{ir} respectively. The periods of the other $m¡r$ servers are all set to zeros. CA then constructs a Lagrange interpolating polynomial function $f_i(X)$ for U_i as

$$f_i(X) = \sum_{j=1}^{m} (U\_ID_i + E\_T_{ij}) \frac{(X - U\_ID_i)}{(S\_SK_j - U\_ID_i)} \times$$

$$\prod_{k=1, k \ne j}^{m} \frac{(X - S\_SK_k)}{(S\_SK_j - S\_SK_k)} +$$

$$U\_R_i \prod_{y=1}^{m} \frac{(X - S\_SK_y)}{(U\_ID_i - S\_SK_y)}$$

$$= a_m X^m + a_{m-1} X^{m-1} + \ldots + a_1 X + a_0 \pmod N$$

- CA stores $f_i(X)$, U_i¡s identity U_ID_i, his two secret keys U_S_i, U_R_i, and one-way function $h(X, Y)$ in smart card U_SC_i. Then, CA sends the card to U_i via a secure channel.

c) The Login Stage: In this phase, when a registered user U_i wants to login server S_j ($1 \le j \le m$), he inserts his smart card U_SC_i to the reader and keys in his U_PW_i. Then, U_SC_i performs the following steps on behalf of U_i:

- U_SC_i gets timestamp t. Then, it generates a secret random number r_1 and computes

$$C_1 = g^{e * r_1} \pmod{N},$$

$$C_2 = (U\_S_1)^{U\_PW_i} \cdot g^{r_1 * h(C_1, t)}$$

$$= g^{U\_PW_i * r_{ui} * d} \cdot g^{r_1 * h(C_1, t)} \pmod{N}, \text{ and}$$

$$P = (S\_ID_j)^{e * r_1} = (g^{S\_SK_j})^{e * r_1} = g^{S\_SK_j * e * r_1} \pmod{N}$$

- Given 1, 2, ¡, m, and P, U_SC_i computes $f_i(1)$, $f_i(2)$, ¡, $f_i(m)$, and $f_i(P)$. Then, it constructs an authentication message $M = \{U\_ID_i, t, C_1, C_2, f_i(1), f_i(2), ¡, f_i(m), f_i(P)\}$ and sends it to S_j, one of the m servers for, $1 \leq j \leq m$.

d) The Server Authentication Stage: In this phase, after receiving the authentication message from U_i, S_j gets current timestamp t_{now} and performs the following steps to verify the login message from U_i:

- S_j checks U_i's identity U_ID_i and determines if t_{now} ¡ t > ΔT. If either of the two checks does nothold, S_j rejects U_i's login message. Otherwise, it continues.

- S_j uses value C_1 and its secret key S_SK_j to derive the value P shown as below.

$$P = (C_1)^{S\_SK_j} \pmod{N}$$

$$= (g^{e * r_1})^{S\_SK_j} \pmod{N}$$

$$= g^{e * r_1 * S\_SK_j} \pmod{N}.$$

Then, it uses these $m + 1$ points $\{(1, f_i(1)), (2, f_i(2)), ¡, (m, f_i(m)), (P, f_i(P))\}$ to reconstruct the interpolating polynomial

$$f_i(X) = a_m X^m + a_{m-1} X^{m-1} + \ldots + a_1 X + a_0 \pmod{N}$$

- He checks to see whether $\dfrac{(C_2)^e}{(C_1)^{h(C_1, t)} \cdot U\_R_i} = 1$. If it holds, user U_i is authentic. Otherwise, S_j rejects U_i's login message.

B. Attack

We show an impersonation attack on Tsaur et al.¡s first protocol. First, an attacker E forges a smart card as follows.

- E enters U_ID_i, randomly chooses a password $U\_PW_i^{(E)}$ and a random number $r_{ui}^{(E)}$, and calculates two secrets:

$$U\_R_i^{(E)} = g^{U\_PW_i^{(E)} * r_{ui}^{(E)} * e} \pmod{N} \text{ and}$$

$$U\_S_i^{(E)} = g^{r_{ui}^{(E)}} \pmod{N}.$$

- Though, E does not know each server¡s private key, he knows these servers¡ identities. Therefore, he uses each server¡s identity to replace the original corresponding private key in polynomial $f_i(X)$ and form another polynomial $f_E(X)$ as shown in following Equation (1).

$$f_E(X) = \sum_{j=1}^{m} (U\_ID_i + E\_T_{ij}) \frac{(X - U\_ID_i)}{(S\_ID_j - U\_ID_i)} \times$$

$$\prod_{k=1, k \neq j}^{m} \frac{(X - S\_ID_k)}{(S\_ID_j - S\_ID_k)} +$$

$$U\_R_i^{(E)} \prod_{y=1}^{m} \frac{(X - S\_ID_y)}{(U\_ID_i - S\_ID_y)} \pmod{N}$$

$$= b_m X^m + b_{m-1} X^{m-1} + \ldots + b_1 X + b_0 \pmod{N}.$$

In login stage, E performs the follows steps:

- E gets timestamp t. Then, he generates a secret random number $r_1^{(E)}$ and computes $C_1^{(E)}$, $C_2^{(E)}$, and $P^{(E)}$ as

$$C_1^{(E)} = g^{e * r_1^{(E)}} \pmod{N},$$

$$C_2^{(E)} = (U\_S_1)^{U\_PW_i^{(E)}} \cdot g^{r_1^{(E)} * h(C_1^{(E)}, t)} \pmod{N},$$

$$P^{(E)} = (S\_ID_j)^{e * r_1^{(E)}} = (g^{S\_SK_j})^{e * r_1^{(E)}}$$

$$= g^{S\_SK_j * e * r_1^{(E)}} \pmod{N}.$$

- Then, E computes $f_E(1), f_E(2), ¡, f_E(m)$, and $f_E(P^{(E)})$ and sends message $M^{(E)} = \{U\_ID_i, t, C_1^{(E)}, C_2^{(E)}, f_E(1), f_E(2), ¡, f_E(m), f_E(P^{(E)})\}$ to server S_j, one of the m servers for $1 \leq j \leq m$.

When receiving message $M^{(E)}$, S_j gets current timestamp t_{now}. It then performs the following verification steps to authenticate E.

- S_j checks E's identity U_ID_i and determines whether t_{now} ¡ $t < \Delta T$. If either of the two checks dose not hold, S_j rejects. Otherwise, he continues.

- S_j uses the transmitted value $C_1^{(E)}$ and his secret key S_SK_j to derive the value $P^{(E)}$, as shown in the following equation, Equation (2).

$$P^{(E)} = (C_1^{(E)})^{S\_SK_j} = (g^{e * r_1^{(E)}})^{S\_SK_j} \pmod{N}$$

$$= g^{e * r_1^{(E)} * S\_SK_j} \pmod{N} \quad ¡ \ ¡ \ ¡ \quad \text{Equation (2)}$$

Then, it uses these $m + 1$ points $\{(1, f_E(1)), (2, f_E(2)), ¡, (m, f_E(m)), (P^{(E)}, f_E(P^{(E)}))\}$ to reconstruct the interpolating polynomial

$$f_E(X) = b_m X^m + b_{m-1} X^{m-1} + \ldots + b_1 X + b_0 \pmod{N}$$

- S_j verifies whether $\dfrac{(C_2^{(E)})^e}{(C_1^{(E)})^{h(C_1^{(E)}, t)} \cdot U\_R_i^{(E)}} = 1$. If it holds, E is authentic.

Obviously, E can pretend as U_i successfully since the computation result is equal to 1, as shown in Equation (3).

$$\frac{(C_2{}^{(E)})^e}{(C_1{}^{(E)})^{h(C_1{}^{(E)},t)} \cdot U\_R_i{}^{(E)}}$$

$$= \frac{(g^{U\_PW_i{}^{(E)}*r_{ui}{}^{(E)}} \cdot g^{r_1{}'*h(C_1{}^{(E)},t)})^e}{g^{e*r_1{}^{(E)}*h(C_1{}^{(E)},t)} \cdot g^{U\_PW_i{}^{(E)}*r_{ui}{}^{(E)}*e}}$$

$$= \frac{g^{U\_PW_i{}^{(E)}*r_{ui}{}^{(E)}*e} \cdot g^{r_1{}^{(E)}*h(C_1{}^{(E)},t)*e}}{g^{e*r_1{}^{(E)}*h(C_1{}^{(E)},t)} \cdot g^{U\_PW_i{}^{(E)}*r_{ui}{}^{(E)}*e}}$$

$$= 1 \pmod{N} \text{ ¡¡¡¡¡¡¡¡¡} \qquad \text{Equation (3)}$$

III. REVIEW AND ATTACK ON TSAUR ET AL.¡S SECOND PROTOCOL

A. Review

Tsaur et al.¡s second protocol [9] consists of four stages. They are (1) The system setup stage, (2) The user registration stage, (3) The login stage, and (4) The server authentication stage. We show them as follows.

1) The System Setup Stage: CA selects a large number p, and publishes a generator g of Z_P^* and an one-way hash function $h(X, Y)$. CA also selects a secret key S_SK_j for server S_j and computes S_j¡s public identity as $S\_ID_j = g^{S\_SK_j} \pmod p$, $1 \le j \le m$.

2) The User Registration Stage: When a new user U_i wants to register at m servers, S_1, S_2, ¡ and S_m (in a multi-server system), he and CA together perform the registration process through a secure channel described as follows:

- U_i chooses his identity U_ID_i and password U_PW_i, and transmits them to CA.

- CA randomly chooses a number r and computes two secret keys:

  $U\_R_i = g^r \pmod p$ and

  $U\_S_i = r^{-U\_PW_i} \pmod p$.

- CA supposes that U_i wants to obtain the services of r servers, S_1, S_2, ¡, and S_r. Assume that the service periods of r servers are E_T_{i1}, E_T_{i2}, ¡, and E_T_{ir} respectively. The periods of the other servers S_{r+1}, S_{r+2}, ¡, and S_m are all set to zeros. CA then uses S_j¡s secret key S_SK_j to construct a Lagrange interpolating polynomial function $f_i(X)$ for U_i as follows:

$$f_i(X) = \sum_{j=1}^{m} (U\_ID_i + E\_T_{ij}) \frac{(X - U\_ID_i)}{(S\_SK_j - U\_ID_i)} \times$$

$$\prod_{k=1,k \ne j}^{m} \frac{(X - S\_SK_k)}{(S\_SK_j - S\_SK_k)} +$$

$$U\_R_i \prod_{y=1}^{m} \frac{(X - S\_SK_y)}{(U\_ID_i - S\_SK_y)}$$

$$= a_m X^m + a_{m-1} X^{m-1} + \ldots + a_1 X + a_0 \pmod p .$$

- CA then stores U_S_i and $f_i(X)$ into the storage of smart card U_SC_i, and sends the card to U_i via a secure channel.

3) The Login Stage: When a registered user U_i wants to login to server S_j, he inserts his smart card U_SC_i to the reader and keys in his password U_PW_i. Then, U_SC_i performs the following steps on behalf of U_i:

- U_SC_i gets timestamp t and computes $r = (U\_S_i)^{U\_PW_i}$. Then, it generates a secret random number r_1 and computes C_1, C_2 and P as

  $C_1 = g^{r_1} \pmod p$,

  $C_2 = r_1 + r \cdot h(C_1, t) \pmod p$, and

  $P = (S\_ID_j)^{r_1} \pmod p$.

- Given $1, 2, ¡, m$, and P, U_SC_i computes $f_i(1), f_i(2), ¡, f_i(m)$, and $f_i(P)$. Then, it constructs message $M = \{U\_ID_i, t, C_1, C_2, f_i(1), f_i(2), ¡, f_i(m), f_i(P)\}$ and sends it to S_j.

4) The Server Authentication Stage: When receiving the authentication message from U_i, S_j obtains current timestamp t_{now} and performs the following steps to verify U_i¡s login message:

- S_j checks U_i's identity U_ID_i and determines whether $t_{now} \ ¡ \ t < \Delta T$. If both hold, S_j computes $P = (C_1)^{S\_SK_j} \pmod p$.

- S_j uses the $m + 1$ points $\{(1, f_i(1)), (2, f_i(2)), ¡, (m, f_i(m)), (P, f_i(P))\}$ from U_ID_i to reconstruct the interpolating polynomial

$$f_i(X) = a_m X^m + a_{m-1} X^{m-1} + \ldots + a_1 X + a_0 \pmod N$$

- S_j checks to see whether $\dfrac{g^{C_2}}{(C_1) \cdot (U\_R_i)^{h(C_1,t)}} = 1$. If it holds, user U_i is authentic. Otherwise, U_i is rejected.

B. Attack

We show an impersonation attack on Tsaur et al.¡s second protocol. First, an attacker E forges a smart card as follows.

- E enters U_ID_i, randomly chooses a password $U\_PW_i^{(E)}$ and a number $r^{(E)}$, and computes two secrets as

  $U\_R_i^{(E)} = g^{r^{(E)}} \pmod p$ and

  $U\_S_i^{(E)} = r^{-U\_PW_i^{(E)}} \pmod p$.

- Though, E does not know each server¡s private key, he knows these servers¡ identities. Therefore, he uses each server¡s identity to replace the original corresponding private key in polynomial $f_i(X)$ and form another polynomial $f_E(X)$ as shown in following Equation (4).

$$f_E(X) = \sum_{j=1}^{m} (U\_ID_i + E\_T_{ij}) \frac{(X - U\_ID_i)}{(S\_ID_j - U\_ID_i)} \times$$

$$\prod_{k=1, k \neq j}^{m} \frac{(X - S\_ID_k)}{(S\_ID_j - S\_ID_k)} +$$

$$U\_R_i^{(E)} \prod_{y=1}^{m} \frac{(X - S\_ID_y)}{(U\_ID_i - S\_ID_y)}$$

$$= b_m X^m + b_{m-1} X^{m-1} + \ldots + b_1 X + b_0 \pmod{p}$$

¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ Equation (4)

In the login stage, when E wants to login to server S_j, he performs the following steps:

- E gets timestamp t and computes $r^{(E)} = (U\_S_i^{(E)})^{U\_PW_i^{(E)}}$. Then, it generates a secret random number $r_1^{(E)}$ and computes $C_1^{(E)}$, $C_2^{(E)}$ and $P^{(E)}$ as

$$C_1^{(E)} = g^{r_1^{(E)}} \pmod{p},$$

$$C_2^{(E)} = r_1^{(E)} + r^{(E)} \cdot h(C_1^{(E)}, t) \pmod{p}, \text{ and}$$

$$P^{(E)} = (S\_ID_j)^{r_1^{(E)}} \pmod{p}.$$

- E computes $f_E(1)$, $f_E(2)$, ¡, $f_E(m)$, and $f_E(P^{(E)})$ and sends message $M^{(E)} = \{U\_ID_i, t, C_1^{(E)}, C_2^{(E)}, f_E(1), f_E(2), ¡, f_E(m), f_E(P^{(E)})\}$ to the server S_j.

After receiving message $M^{(E)}$, S_j gets current timestamp t_{now}. He then performs the following verification steps to authenticate E.

- S_j checks E's identity U_ID_i and determines whether t_{now} ¡ $t < \Delta T$. If both hold, S_j computes

$$P^{(E)} = (C_1^{(E)})^{S\_SK_j} \pmod{p}.$$

- S_j uses the $m + 1$ points $\{(1, f_E(1)), (2, f_E(2)), ¡, (m, f_E(m)), (P, f_E(P))\}$ to reconstruct the interpolating polynomial

$$f_E(X) = b_m X^m + b_{m-1} X^{m-1} + \ldots + b_1 X + b_0 \pmod{p}$$

- S_j verifies if $\dfrac{g^{C_2^{(E)}}}{(C_1^{(E)}) \cdot (U\_R_i^{(E)})^{h(C_1^{(E)}, t)}} = 1$. If it holds, E is authentic.

Obviously, E can pretend as U_i successfully. Since that the computation result of the verification is obviously equal to 1, as shown in following Equation (5).

$$\frac{g^{C_2^{(E)}}}{(C_1^{(E)}) \cdot (U\_R_i^{(E)})^{h(C_1^{(E)}, t)}}$$

$$= \frac{g^{r_1^{(E)} + r^{(E)} * h(C_1^{9E}, t)}}{g^{r_1^{(E)}} \cdot g^{r^{(E)} * h(C_1^{(E)}, t)}}$$

$$= \frac{g^{r_1^{(E)} + r^{(E)} * h(C_1^{(E)}, t)}}{g^{r_1^{(E)} + r^{(E)} * h(C_1^{(E)}, t)}}$$

$$= 1 \pmod{p} \quad \text{¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡ ¡} \qquad \text{Equation (5)}$$

IV. CONCLUSION

In this paper, we present the security analyses of Tsaur et al.¡s two smart card based password authentication protocols in multi-server environments. Our results show that they are both vulnerable and suffer from the impersonation attacks which we have described in this article.

REFERENCES

[1] A. K. Awasthi and S. Lal ¡An enhanced remote user authentication scheme using smart cards, ¡*IEEE Trans. Consumer Electron.,* vol. 50, No. 2, pp. 583-586, May 2004.

[2] C.K. Chan, L.M. Cheng, ¡Cryptanalysis of a remote user authentication scheme using smarts cards, ¡*IEEE Transactions on Consumer Electronics,* vol. 46, no 4, pp. 992¡ 993, 2000.

[3] C.C. Chang, T.C. Wu, ¡Remote password authentication scheme with smart cards, ¡*IEE Proceedings-Computers and Digital Techniques,* vol. 138, issue 3, pp.165¡ 168, 1991.

[4] T. Hwang, Y. Chen, and C.S. Laih, ¡Non-interactive password authentications without password tables, ¡*IEEE Region 10 Conference on Computer and Communication Systems, IEEE Computer Society,* Vol. 1, pp.429¡ 431, 1990.

[5] M.S. Hwang and L.H. Li, ¡A new remote user authentication scheme using smart cards, ¡*IEEE Transactions on Consumer Electronics,* vol.46, no.1, pp. 28-30, Feb. 2000.

[6] K. C. Leung, L. M. Cheng, A. S. Fong and C. K. Chan, ¡Cryptanalysis of a modified remote user authentication scheme using smart cards, ¡*IEEE Trans. Consumer Electron.,* vol. 49, No. 4, pp. 1243-1245, Nov. 2003.

[7] J.J. Shen, C. W. Lin and M. S. Hwang, ¡A modified remote user authentication scheme using smart cards, ¡*IEEE Trans. Consumer Electron.,* vol. 49, No. 2, pp. 414-416, May 2003.

[8] W.J. Tsaur, C.C. Wu, W.B. Lee, ¡A smart card-based remote scheme for password authentication in multi-server Internet services, ¡*Computer Standards & Interfaces,* Vol. 27, No. 1, pp. 39-51, November 2004.

[9] W.J. Tsaur, C.C. Wu, W.B. Lee, ¡An enhanced user authentication scheme for multi-server Internet services, ¡*Applied Mathematics and Computation,* Vol. 170, No. 1-1, pp. 258-266, November 2005.

[10] G. Yang, D. S. Wong, H. Wang, X. Deng, ¡Two-factor mutual authentication based on smart cards and passwords¡, *Journal of Computer and System Sciences,* Vol. 74, No. 7, pp.1160-1172, November 2008.

[11] T. Goriparthi, M. L. Das, A. Saxena, ¡An improved bilinear pairing based remote user authentication scheme¡, *Computer Standards & Interfaces,* Vol. 31, No. 1, pp. 181-185, January 2009.

[12] I. E. Liao, C. C. Lee, M. S. Hwang, ¡A password authentication scheme over insecure networks¡, *Journal of Computer and System Sciences,* Vol. 72, No. 4, pp. 727-740, June 2006.

[13] J. Y. Liu, A. M. Zhou, M. X. Gao, ¡A new mutual authentication scheme based on nonce and smart cards¡, Computer Communications, Vol. 31, No. 10, pp. 2205-2209, June 2008.

[14] H. S. Rhee, J. O. Kwon, D. H. Lee, ¡A remote user authentication scheme without using smart cards¡, *Computer Standards & Interfaces,* Vol. 31, No. 1, pp. 6-13, January 2009.

[15] S. K. Kim , M. G. Chung, ¡More secure remote user authentication scheme¡, *Computer Communications,* Vol. 32, No. 6, pp. 1018-1021, April 2009.

[16] Y. Y. Wang, J. Y. Liu, F. X. Xiao, J. Dan, ¡A more efficient and secure dynamic ID-based remote user authentication scheme¡, *Computer Communications*, Vol. 32, No. 4, pp. 583-585, March 2009.

[17] J. Xu, W. T. Zhu, D. G. Feng, ¡An improved smart card based password authentication scheme with provable security¡, *Computer Standards & Interfaces*, Vol. 31, No. 4, pp. 723-728, June 2009.

[18] M. S. Hwang, S. K. Chong, T. Y. Chen, ¡DoS-resistant ID-based password authentication scheme using smart cards¡, *Journal of Systems and Software*, Vol. 83, No. 1, pp. 163-172, January 2010.

[19] C. T. Li, M. S. Hwang, ¡An efficient biometrics-based remote user authentication scheme using smart cards¡, *Journal of Network and Computer Applications*, Vol. 33, No. 1, pp. 1-5, January 2010.

[20] D. Z. Sun, J. P. Huai, J. Z. Sun, J. X. Li, J. W. Zhang, Z. Y. Feng, ¡Improvements of Juang et al.¡s Password-Authenticated Key Agreement Scheme Using Smart Cards¡, *IEEE Transactions on Industrial Electronics*, Vol. 56, No. 6, pp. 2284-2291, June 2009.

AUTHORS PROFILE

Jue-Sam Chou received his Ph.D. degree in the department of computer science and information engineering from National Chiao Tung Univ. (NCTU) in Hsinchu, Taiwan,ROC. He is an associate professor and teaches at the department of Info. Management of Nanhua Univ. in Chiayi, Taiwan. His primary research interests are electronic commerce, data security and privacy, protocol security, authentication, key agreement, cryptographic protocols, E-commerce protocols, and so on.

Chun-Hui Huang is now a graduate student at the department of Info. Management of Nanhua Univ. in Chiayi, Taiwan. She is also a teacher at Nantou County Shuang Long Elementary School in Nantou, Taiwan. Her primary interests are data security and privacy, protocol security, authentication, key agreement.

Yalin Chen received her bachelor degree in the depart. of computer science and information engineering from Tamkang Univ. in Taipei, Taiwan and her MBA degree in the department of information management from National Sun-Yat-Sen Univ. (NYSU) in Kaohsiung, Taiwan. She is now a Ph.D. candidate of the Institute of Info. Systems and Applications of National Tsing-Hua Univ.(NTHU) in Hsinchu, Taiwan. Her primary research interests are data security and privacy, protocol security, authentication, key agreement, electronic commerce, and wireless communication security.

An Efficient Feature Extraction Technique
for Texture Learning

R. Suguna

Research Scholar, Department of Information Technology
Madras Institute of Technology, Anna University
Chennai- 600 044, Tamil Nadu, India.
hitec_suguna@hotmail.com

P. Anandhakumar

Assistant Professor, Department of Information Tech.
Madras Institute of Technology, Anna University
Chennai- 600 044, Tamil Nadu, India.
anandh@annauniv.edu

Abstract— **This paper presents a new methodology for discovering features of texture images. Orthonormal Polynomial based Transform is used to extract the features from the images. Using orthonormal polynomial basis function polynomial operators with different sizes are generated. These operators are applied over the images to capture the texture features. The training images are segmented with fixed size blocks and features are extracted from it. The operators are applied over the block and their inner product yields the transform coefficients. These set of transform coefficients form a feature set of a particular texture class. Using clustering technique, a codebook is generated for each class. Then significant class representative vectors are calculated which characterizes the textures. Once the orthonormal basis function of particular size is found, the operators can be realized with few matrix operations and hence the approach is computationally simple. Euclidean Distance measure is used in the classification phase. The transform coefficients have rotation invariant capability. In the training phase the classifier is trained with samples with one particular angle of image and tested with samples at different angles. Texture images are collected from Brodatz album. Experimental results prove that the proposed approach provides good discrimination between the textures.**

Keywords- Texture Analysis; Orthonormal Transform; codebook generation; Texture Class representatives; Texture Characterization.

I. INTRODUCTION

Texture can be regarded as the visual appearance of a surface or material. Textures appear in numerous objects and environments in the universe and they can consist of very different elements. Texture analysis is a basic issue in image processing and computer vision. It is a key problem in many application areas, such as object recognition, remote sensing, content-based image retrieval and so on. A human may describe textured surfaces with adjectives like fine, coarse, smooth or regular. But finding the correlation with mathematical features indicating the same properties is very difficult. We recognize texture when we see it but it is very difficult to define. In computer vision, the visual appearance of the view is captured with digital imaging and stored as image

pixels. Texture analysis researchers agree that there is significant variation in intensity levels or colors between nearby pixels and at the limit of resolution there is non-homogeneity. Spatial non-homogeneity of pixels corresponds to the visual texture of the imaged material which may result from physical surface properties such as roughness, for example. Image resolution is important in texture perception, and low-resolution images contain typically very homogenous textures.

The appearance of texture depend upon three ingredients: (i) some local 'order' is repeated over a region which is large in comparison to the order's size, (ii) the order consists in the nonrandom arrangement of elementary parts, and (iii) the parts are roughly uniform entities having approximately the same dimensions everywhere within the textured region[1].

Image texture, defined as a function of the spatial variation in pixel intensities (gray values), is useful in a variety of applications and has been a subject of intense study by many researchers. One immediate application of image texture is the recognition of image regions using texture properties. Texture is the most important visual cue in identifying these types of homogeneous regions. This is called *texture classification*. The goal of texture classification then is to produce a classification map of the input image where each uniform textured region is identified with the texture class it belongs to [2].

Texture analysis methods have been utilized in a variety of application domains. Texture plays an important role in automated inspection, medical image processing, document processing and remote sensing. In the detection of defects in texture images, most applications have been in the domain of textile inspection. Some diseases, such as interstitial fibrosis, affect the lungs in such a manner that the resulting changes in the X-ray images are texture changes as opposed to clearly delineated lesions. In such applications, texture analysis methods are ideally suited for these images. Texture plays a significant role in document processing and character recognition. The text regions in a document are characterized by their high frequency content. Texture analysis has been extensively used to classify remotely sensed images. Land use classification where homogeneous regions with different types of terrains (such as wheat, bodies of water, urban regions, etc.) need to be identified is an important

application. Haralick et al. [3] used gray level co-occurrence features to analyze remotely sensed images.

Since we are interested in interpretation of images we can define texture as the characteristic variation in intensity of a region of an image which should allow us to recognize and describe it and outline its boundaries. The degrees of randomness and of regularity will be the key measure when characterizing a texture. In texture analysis the similar textural elements that are replicated over a region of the image are called texels. This factor leads us to characterize textures in the following ways:

- The texels will have various sizes and degrees of uniformity

- The texels will be oriented in various directions

- The texels will be spaced at varying distances in different directions

- The contrast will have various magnitudes and variations

- Various amounts of background may be visible between texels

- The variations composing the texture may each have varying degrees of regularity

It is quite clear that a texture is a complicated entity to measure. The reason is primarily that many parameters are likely to be required to characterize it. Characterization of textured materials is usually very difficult and the goal of characterization depends on the application. In general, the aim is to give a description of analyzed material, which can be, for example, the classification result for a finite number of classes or visual exposition of the surfaces. It gives additional information compared only to color or shape measurements of the objects. Sometimes it is not even possible to obtain color information at all, as in night vision with infrared cameras. Color measurements are usually more sensitive to varying illumination conditions than texture, making them harder to use in demanding environments like outdoor conditions. Therefore texture measures can be very useful in many real-world applications, including, for example, outdoor scene image analysis.

To exploit texture in applications, the measures should be accurate in detecting different texture structures, but still be invariant or robust with varying conditions that affect the texture appearance. Computational complexity should not be too high to preserve realistic use of the methods. Different applications set various requirements on the texture analysis methods, and usually selection of measures is done with respect to the specific application.

Typically textures and the analysis methods related to them are divided into two main categories with different computational approaches: the stochastic and the structural methods. Structural textures are often man-made with a very regular appearance consisting, for example, of line or square primitive patterns that are systematically located on the surface (e.g. brick walls). In structural texture analysis the properties and the appearance of the textures are described with different rules that specify what kind of primitive elements there are in

the surface and how they are located. Stochastic textures are usually natural and consist of randomly distributed texture elements, which again can be, for example, lines or curves (e.g. tree bark). The analysis of these kinds of textures is based on statistical properties of image pixels and regions. The above categorization of textures is not the only possible one; there exist several others as well, for example, artificial vs. natural or micro textures vs. macro textures. Regardless of the categorization, texture analysis methods try to describe the properties of the textures in a proper way. It depends on the applications what kind of properties should be sought from the textures under inspection and how to do that. This is rarely an easy task.

One of the major problems when developing texture measures is to include invariant properties in the features. It is very common in a real-world environment that, for example, the illumination changes over time, and causes variations in the texture appearance. Texture primitives can also rotate and locate in many different ways, which also causes problems. On the other hand, if the features are too invariant, they might not be discriminative enough.

II. TEXTURE MODELS

Image texture has a number of perceived qualities which play an important role in describing texture. One of the defining qualities of texture is the spatial distribution of gray values. The use of statistical features is therefore one of the early methods proposed in the machine vision literature.

The gray-level co-occurrence matrix approach is based on studies of the statistics of pixel intensity distributions. The early paper by Haralick et al.[4] presented 14 texture measures and these were used successfully for classification of many types of materials for example, wood, corn, grass and water. However, Conners and Harlow [5] found that only five of these measures were normally used, viz. "energy", "entropy", "correlation", "local homogeneity", and "inertia". The size of the co-occurrence matrix is high and suitable choice of d (distance) and θ (angle) has to be made to get relevant features.

A novel texture energy approach is presented by Laws [6]. This involved the application of simple filters to digital images. The basic filters he used were common Gaussian, edge detector, and Laplacian-type filters and were designed to highlight points of high "texture energy" in the image. Ade investigated the theory underlying Laws' approach and developed a revised rationale in terms of Eigen filters [7]. Each eigenvalue gives the part of the variance of the original image that can be extracted by the corresponding filter. The filters that give rise to low variances can be taken to be relatively unimportant for texture recognition.

The structural models of texture assume that textures are composed of texture primitives. The texture is produced by the placement of these primitives according to certain placement rules. This class of algorithms, in general, is limited in power unless one is dealing with very regular textures. Structural texture analysis consists of two major steps: (a) extraction of the texture elements, and (b) inference of the placement rule. An approach to model the texture by

structural means is described by Fu [8]. In this approach the texture image is regarded as texture primitives arranged according to a placement rule. The primitive can be as simple as a single pixel that can take a gray value, but it is usually a collection of pixels. The placement rule is defined by a tree grammar. A texture is then viewed as a string in the language defined by the grammar whose terminal symbols are the texture primitives. An advantage of this method is that it can be used for texture generation as well as texture analysis.

Model based texture analysis methods are based on the construction of an image model that can be used not only to describe texture, but also to synthesize it. The model parameters capture the essential perceived qualities of texture. Markov random fields (MRFs) have been popular for modeling images. They are able to capture the local (spatial) contextual information in an image. These models assume that the intensity at each pixel in the image depends on the intensities of only the neighboring pixels. Many natural surfaces have a statistical quality of roughness and self-similarity at different scales. Fractals are very useful and have become popular in modeling these properties in image processing.

However, the majority of existing texture analysis methods makes the explicit or implicit assumption that texture images are acquired from the same viewpoint (e.g. the same scale and orientation). This gives a limitation of these methods. In many practical applications, it is very difficult or impossible to ensure that images captured have the same translations, rotations or scaling between each other. Texture analysis should be ideally invariant to viewpoints. Furthermore, based on the cognitive theory and our own perceptive experience, given a texture image, no matter how it is changed under translation, rotation and scaling or even perspective distortion, it is always perceived as the same texture image by a human observer. Invariant texture analysis is thus highly desirable from both the practical and theoretical viewpoint.

Recent developments include the work with automated visual inspection in work. Ojala et al., [9] and Manthalkar et al., [10] aimed at rotation invariant texture classification. Pun and Lee [11] aims at scale invariance. Davis [12] describes a new tool (called polarogram) for image texture analysis and used it to get invariant texture features. In Davis's method, the co-occurrence matrix of a texture image must be computed prior to the polarograms. However, it is well known that a texture image can produce a set of co-occurrence matrices due to the different values of a and d. This also results in a set of polarograms corresponding to a texture. Only one polarogram is not enough to describe a texture image. How many polarograms are required to describe a texture image remains an open problem. The polar grid is also used by Mayorga and Ludeman [13] for rotation invariant texture analysis. The features are extracted on the texture edge statistics obtained through directional derivatives among circularly layered data. Two sets of invariant features are used for texture classification. The first set is obtained by computing the circularly averaged differences in the gray level between pixels. The second computes the correlation function along circular levels. It is demonstrated by many recent publications that Zernike moments perform well in practice to obtain geometric invariance.

Local frequency analysis has been used for texture analysis. One of the best known methods uses Gabor filters and is based on the magnitude information [14]. Phase information has been used in [15] and histograms together with spectral information in [16]. Ojala T & Pietikäinen M [17] proposed a multichannel approach to texture description by approximating joint occurrences of multiple features with marginal distributions, as 1-D histograms, and combining similarity scores for 1-D histograms into an aggregate similarity score. Ojala T introduced a generalized approach to the gray scale and rotation invariant texture classification method based on local binary patterns [18]. The current status of a new initiative aimed at developing a versatile framework and image database for empirical evaluation of texture analysis algorithms is presented by him. Another frequently used approach in texture description is using distributions of quantized filter responses to characterize the texture (Leung and Malik), (Varma and Zisserman) [19] [20]. Ahonen T, proved that the local binary pattern operator can be seen as a filter operator based on local derivative filters at different orientations and a special vector quantization function [21].

A rotation invariant extension to the blur insensitive local phase quantization texture descriptor is presented by Ojansivu V [22].

Unitary Transformations are also used to represent the images. The simple and powerful class of transform coding is linear block transform coding, where the entire image is partitioned into a number of non-overlapping blocks and then the transformation is applied to yield transform coefficients. This is necessitated because of the fact that the original pixel values of the image are highly correlated. A framework using orthogonal polynomials for edge detection and texture analysis is presented in [23] [24].

III. ORTHONORMAL POLYNOMIAL TRANSFORM

A linear 2-D image formation system usually considered around a Cartesian coordinate separable, blurring, point spread operator in which the image *I* results in the superposition of the point source of impulse weighted by the value of the object *f*. Expressing the object function *f* in terms of derivatives of the image function *I* relative to its Cartesian coordinates is very useful for analyzing the image. The point spread function M(x, y) can be considered to be real valued function defined for (x, y) $\in$ X x Y, where X and Y are ordered subsets of real values. In case of gray-level image of size (n x n) where X (rows) consists of a finite set, which for convenience labeled as {0, 1, 2, … ,n-1}, the function M(x, y) reduces to a sequence of functions.

$$M(i,t) = u_i(\tau), \ \tau=0,1,...\nu-1 \qquad (1)$$

The linear two dimensional can be defined by the point spread operator M(x,y), (M(i,t) = u_i(t)) as shown in equation 2.

$$\beta'(\zeta,\eta) = \int_{x\in X}\int_{y\in Y} M(\zeta,x)M(\eta,y)I(x,y)dxdy \qquad (2)$$

Considering both X and Y to be a finite set of values {0, 1, 2, …. ,n-1}, equation (2) can be written in matrix notation as follows

$$|\beta'_{ij}| = (|M| \otimes |M|)'|I| \qquad (3)$$

where $\otimes$ is the outer product, $|\beta'_{ij}|$ are n^2 matrices arranged in the dictionary sequence, $|I|$ is the image, $|\beta'_{ij}|$ are the coefficients of transformation and the point spread operator $|M|$ is

$$|M| = \begin{vmatrix} u_0(t_1) & u_1(t_1) \dots & u_{n-1}(t_1) \\ u_0(t_2) & u_1(t_2) \dots & u_{n-1}(t_2) \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ u_0(t_n) & u_0(t_n) \dots & u_{n-1}(t_n) \end{vmatrix} \qquad (4)$$

We consider the set of orthogonal polynomials $u_0(t), u_1(t), \dots, u_{n-1}(t)$ of degrees 0, 1, 2, …, n-1 respectively to construct the polynomial operators of different sizes from equation (4) for n ≥ 2 and $t_i = i$. The generating formula for the polynomials is as follows.

$$u_{i+1}(t) = (t - \mu)u_i(t) - b_i(n)u_{i-1}(t) \; for \; i \geq 1 \qquad (5)$$

$$u_1(t) = t - \mu, and \; u_0(t) = 1,$$

where

$$b_i(n) = \frac{\langle u_i, u_i \rangle}{\langle u_{i-1}, u_{i-1} \rangle} = \frac{\sum_{t=1}^n u_i{}^2(t)}{\sum_{t=1}^n u_{i-1}{}^2(t)} \qquad (6)$$

and

$$\mu = \frac{1}{n} \sum_{t=1}^n t \qquad (7)$$

Considering the range of values of t to be $t_i = i$, $i = 1,2,3,\dots n$, we get

$$b_i(n) = \frac{i^2(n^2 - i^2)}{4(4i^2 - 1)} \qquad (8)$$

$$\mu = \frac{1}{n} \sum_{t=1}^n t = \frac{n+1}{2} \qquad (9)$$

We can construct point-spread operators $|M|$ of different size from equation (4) using the above orthogonal polynomials for $n \geq 2$ and $t_i = i$.

The orthogonal basis functions for n=2 and n=3 are given below.

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & -2 \\ 1 & 1 & 1 \end{bmatrix}$$

Orthonormal basis functions can be derived from orthogonal sets. Suppose that S is a set of vectors in an inner product space.

(a) If each pair of distinct vectors from S is orthogonal then we call S an orthogonal set.

(b) If S is an orthogonal set and each of the vectors in S also has a norm of 1 then we call S an orthonormal set.

To enforce orthonormal property, divide each vector by its norm. Suppose $S = (v_1, v_2, v_3)$ forms an orthogonal set. Then, $\langle v_1, v_2 \rangle = \langle v_2, v_3 \rangle = \langle v_1, v_3 \rangle = 0$. Any vector v can be turned into a vector with norm 1 by dividing by its norm as follows,

$$\frac{1}{\|v\|} \; v \qquad (10)$$

To convert S to have orthonormal property, divide each vector by its norm.

$$u_i = \frac{1}{\|v_i\|} \; v_i \, , i = 1, 2, 3 \qquad (11)$$

After finding the orthonormal basis function, the operators are generated by applying outer product. For an orthonormal basis function of size n, n^2 operators are generated. Applying the operators over the block of the image we get transform coefficients.

IV. METHODOLOGY

Sample Images representing different Textures are collected. We collected the images from Outex Texture Database. Each image is of size 128 x 128. Images of each texture are partitioned into two groups as Training Set and Test Set.

The process involved in capturing the texture characterization is depicted in Figure-1. Each training image is partitioned into non-overlapping blocks of size M*M. We have chosen M = 4. Features are extracted from each block using orthonormal polynomial based transform as described in section 3. From each block a k-dimensional feature vector is generated. A codebook is built for each concept classes. The algorithm for construction of codebook is discussed below.

A. Codebook Generation Algorithm

Input: Training Images of Texture Ti
Output: Codebook of the Texture Ti

1. Read the image Tr(m) from the Texture Class Ti, where m=1,2,…M; M denotes the number of training images in Ti and i=1,2,…L; L denotes the number of Textures. Size of Tr(m) is 128x128.

2. Each image is partitioned into p x p blocks and we have P blocks for each training image, p=4.

3. For each block apply Orthonormal Based transform by using a set of (pxp) polynomial operators and extract the feature coefficients. Inner product between a polynomial operator and image block results in a transform coefficient. We get p² coefficients for each block.

4. Rearrange the feature coefficients into 1-D array in descending sequence.

Figure 1.　　　Process involved in Texture Characterization

5. Take only d coefficients to form the feature vector **z**, where **z** = {z(j), j=1,2,…,d; d<k}.

6. From P blocks get P x d coefficients.

7. Repeat 2-6 for all images in T_i and collect the **z** vectors.

Apply clustering technique, to cluster the feature vectors of T_i. The number clusters decides the codebook size. The mean of the clusters form the code vectors.

B. Building Class Representative Vector

Input:　　　　Images of size N x N, Texture codebook

Output:　　　Class Representative Vector R_i.

1. For each image in T_i, generate the code indices associated with the corresponding codebook.

2. Find the number of occurrences in each code index for each image.

3. Compute the mean of occurrences to generate class representative vector R_i, where i=1,2,…L, where L is the number of Textures.

4. Repeat 1-3 for all T_i.

C. Texture Classification

Given any texture image this phase determines to which texture class the image is relevant. Images from the Test set are partitioned into non-overlapping blocks of size M*M. Features are extracted using orthonormal polynomial Transform. Consulting the codebooks, code indices are generated and the corresponding input representative vector is formed. Compute the distance d_i between the Class Representative vector Ri and input image representative vector IR_i for T_i. Euclidean distance is used for similarity measure.

$$d_i = dist(IR_i, R_i)$$

Find min (d_i) to obtain the Texture class.

V. RESULTS AND DISCUSSION

We demonstrate the performance of our approach with the proposed Transform coefficients with texture image data that have been used in recent studies on rotation invariant texture classification. Since the data included samples from several rotation angles, we also present results for a more challenging setup, where the samples of just one particular rotation angle

are used for training the texture classifier, which is then tested with the samples of the other rotation angles.

A. Image Data and Experimental Setup

The image data included 12 textures from the Brodatz album. Textures are presented at 6 different rotation angles (0, 30, 60, 90, 120, and 150). For each texture class there were 16 images for each class and angle (hence 1248 images in total). Each texture class comprises following subsets of images: 16 'original' images, 16 images rotated at 30^0, 16 images rotated at 60^0, 16 images rotated at 90^0, 16 images rotated at 120^0 and 16 images rotated at 150^0. The size of each image is 128x128.

The texture classes considered for our study are shown in Figure. 2. The texture classes are divided into two sets. Texture Set-1 contains structural textures (regular patterns) and Texture Set-2 contains stochastic textures (irregular patterns).

Texture Set-1 includes {bark, brick, bubbles, raffia, straw, weave}. Texture Set-2 includes {grass, leather, pigskin, sand, water, wool}.

The statistical features of the texture class are studied first. The mean and variance of the texture classes are found and depicted in Figure-3 to Figure 6.
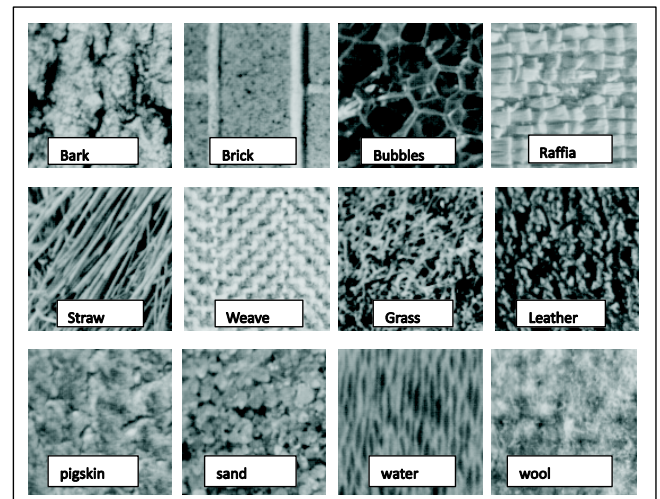
Figure 2.　　　Sample Images of Textures

Mean of Texture - Set1

Figure 3.   Mean of Structural Textures

Mean of Texture - Set2
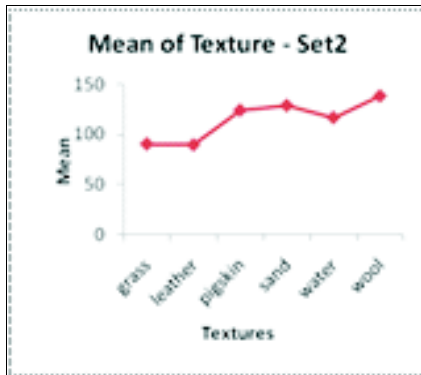
Figure 4.   Mean of Stochastic Textures

Variances of Texture - Set1
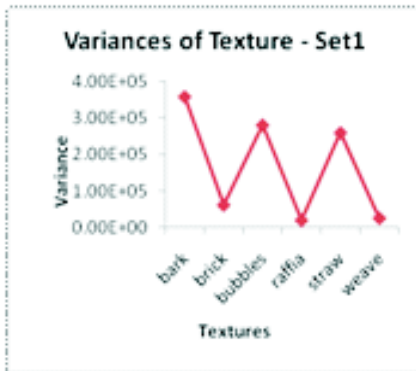
Figure 5.   Variance of Structural Textures

Variances of Texture - Set2
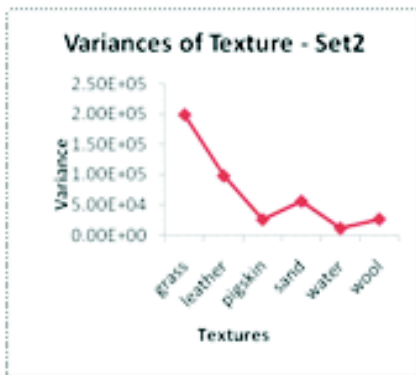
Figure 6.   Variance of Stochastic Textures

B.   Contribution of Transform Coefficients

Each Texture class with rotation angle 0 is taken for training. Other images are used for Testing. For each Texture class a code book is generated with the training samples. A Class Representative Vector is estimated. Figure 7 and Figure 8 shows the representatives of Textures.
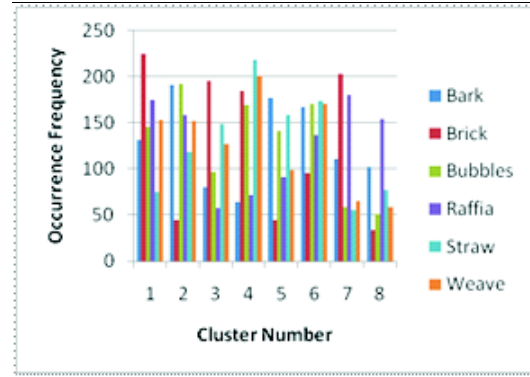
Figure 7.   Class Representatives of Structural Textures
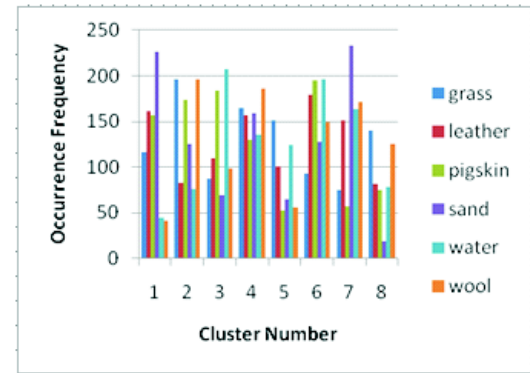
Figure 8.   Class Representatives of Stochastic Textures

Table 1 and Table 2 presents results for a the challenging experimental setup where the classifier is trained with samples of just one rotation angle and tested with samples of other rotation angles.

Texture	Classification Accuracy (%) for different Training angles				
	30^0	60^0	90^0	120^0	150^0
Bark	86.6	68.75	86.6	75.0	68.75
Brick	75.0	86.6	87.5	75.0	86.6
Bubbles	93.75	93.75	100	100	100
Raffia	100	93.75	87.5	87.5	87.5
Straw	56.25	62.5	68.75	56.25	62.5
Weave	93.75	100	100	100	93.75

Table 1 Classification Accuracies (%) of Structural Textures trained with One rotation angle (0^0) and Tested with other versions

Texture	Classification Accuracy (%) for different Training angles				
	30^0	60^0	90^0	120^0	150^0
Grass	100	100	100	93.75	93.75
Leather	87.5	93.75	87.5	93.75	93.75
Pigskin	87.5	87.5	93.75	75.0	68.75
Sand	75.0	75.0	68.75	68.75	75.0
Water	100	93.75	93.75	87.5	87.5
Wool	86.6	86.6	62.5	68.75	75

Table 2 Classification Accuracies (%) of stochastic Textures trained with One rotation angle (0^0) and Tested with other versions

It is observed that in Structural Textures, Bark is misclassified as Straw and few as Brick. Brick is misclassified as Raffia. Straw is misclassified as Bark and Bubbles. In the case of Stochastic Textures Sand is misclassified as pigskin. Wool is misclassified as Pigskin and sand. Compared to structural Textures the performance of stochastic Textures is good. The performance of Structured Textures and Stochastic Textures are shown in Figure 9 and Figure 10.
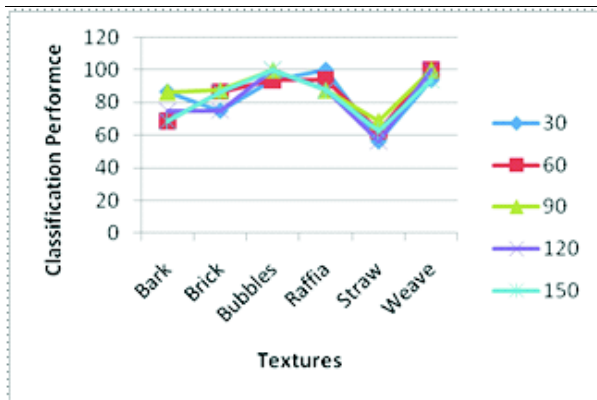
Figure 9.   Classification Performance of Structural Textures
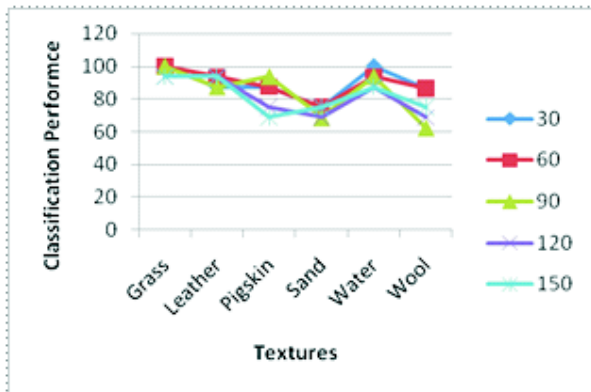
Figure 10.   Classification Performance of Stochastic Textures

The overall performance of Structured and Stochastic Textures is reported in Figure 11 and Figure 12. If the mean difference between the textures is less, then their classification performance degrades.
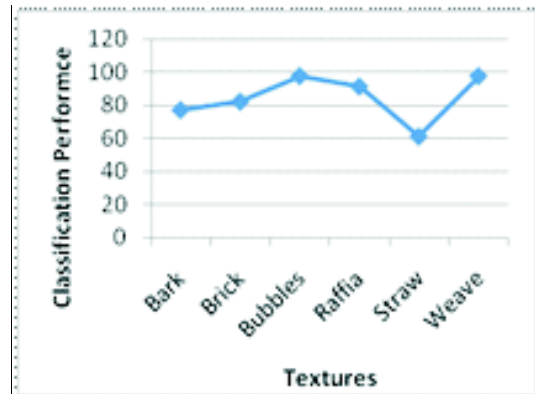
Figure 11.   Overall Classification Performance of Structural Textures
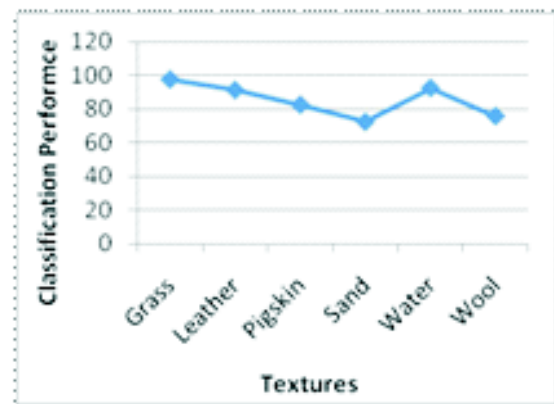
Figure 12.   Overall Classification Performance of Stochastic Textures

We have also compared the performance of our feature extraction method with other approaches. Table 3 shows the comparative study with other Texture models.

Texture model	Recognition rate in %
Co occurrence matrix	78.6
Autocorrelation method	76.1
Laws Texture measure	82.2
Orthonormal Transformed Feature Extraction	89.2

Table 3 Performance of various Texture measures in classication

VI. CONCLUSION

An efficient way of extracting features from textures is presented. From the orthonormal basis, new operators are generated. These operators perform well in characterizing the textures. The operator can be used for gray-scale and rotation invariant texture classification. Experimental results are appreciable where the original version of image samples are used for learning and tested for different rotation angles. Computational simplicity is another advantage since the operator is evaluated by computing the inner product. This facilitates less time for implementation. The efficiency can be further improved by varying the codebook size and the dimension of feature vectors.

REFERENCES

[1] Hawkins J K Textural Properties for Pattern Recognition in Picture Processing *and Psychopictorics*, (LIPKIN B AND ROSENFELD A Eds), Academic Press, New York 1969.

[2] Chen Ch, Pau Lf, Wang Psp (1998) The Handbook Of Pattern Recognition And Computer Vision (2nd Edition), Pp. 207-248, World Scientific Publishing Co., 1998.

[3] Haralick Rm, Shanmugam K And Dinstein I, (1973) Textural Feature For Image Classification Ieee Transactions On Systems, Man, And Cybernetics, Smc-3, Pp. 610-621.

[4] Haralick Rm (1979) Statistical And Structural Approaches To Texture, Proc Ieee 67,No,5, 786-804

[5] Conners Rw And Harlow Ca (1980) Toward A Structural Textural Analyzer Based On Statistical Methods, Comput. Graph, Image Processing,12,224-256.

[6] Laws Ki (1979) Texture Energy Measures Proc. Image Understanding Workshop, Pp. 47-51. 1979.

[7] Ade F (1983) Characterization Of Texture By Eigenfilters Signal Processing, 5, No.5, 451-457.

[8] Fu Ks (1982) Syntactic Pattern Recognition And Applications, Prentice-Hall, New Jersey, 1982.

[9] Ojala T, Pietikainen M And Maenpaa T (2002) Multiresolution Gray-Scale And Rotation Invariant Texture Classification With Local Binary Patterns, Ieee Trans. Pattern Anal. Mach. Intell, 24, No. 7, 971-987.

[10] Manthalkar R, Biswas Pk And Chatterji Bn (2003) Rotation Invariant Texture Classification Using Even Symmetric Gabor Filters, Pattern Recog. Lett, 24, No. 12, 2061-2068.

[11] Pun Cm And Lee Mc (2003) Log Polar Wavelet Energy Signatures For Rotation And Scale Invariant Texture Classification, Ieee Trans. Pattern. Anal, Mach. Intell. 25, No.5, 590-603.

[12] Larry S Davis (1981) Polarogram: A New Tool For Image Texture Analysis, Pattern Recognition 13 (3) 219–223.

[13] Mayorga L. Ludman (1994), Shift And Rotation Invariant Texture Recognition With Neural Nets, Proceedings Of Ieee International Conference On Neural Networks, Pp. 4078–4083.

[14] Manthalkar R, Biswas Pk And Chatterji Bn (2003) Rotation Invariant Texture Classification Using Even Symmetric Gabor Filters, Pattern Recog. Lett, 24, No. 12, 2061-2068.

[15] Vo Ap, Oraintara S, And Nguyen Tt (2007) Using Phase And Magnitude Information Of The Complex Directional Filter Bank For Texture Image Retrieval, Proc. Ieee Int. Conf. On Image Processing (Icip'07), Pages 61–64.

[16] Xiuwen L And Deliang W(2003) Texture Classification Using Spectral Histograms, Ieee Trans. Image Processing, 12(6):661–670.

[17] Ojala T & Pietikäinen M (1998) Nonparametric Multichannel Texture Description With Simple Spatial Operators, Proc. 14th International Conference On Pattern Recognition, Brisbane, Australia, 1052 – 1056.

[18] Ojala T, Pietikäinen M & Mäenpää T (2001) A Generalized Local Binary Pattern Operator For Multiresolution Gray Scale And Rotation Invariant Texture Classification, Advances In Pattern Recognition, Icapr

2001 Proceedings, Lecture Notes In Computer Science 2013, Springer, 397 - 406.

[19] Leung T And Malik J (2001) Representing And Recognizing The Visual Appearance Of Materials Using Three Dimensional Textons, Int. J. Comput. Vision, 43(1):29– 44.

[20] Varma M And Zisserman A (2005) A Statistical Approach To Texture Classification From Single Images, International Journal Of Computer Vision, 62(1–2):61–81.

[21] Ahonen T & Pietikäinen M (2008) A Framework For Analyzing Texture Descriptors", Proc. Third International Conference On Computer Vision Theory And Applications (Visapp 2008), Madeira, Portugal, 1:507-512.

[22] Ojansivu V & Heikkilä J (2008) A Method For Blur And Affine Invariant Object Recognition Using Phase-Only Bispectrum, Proc. Image Analysis And Recognition (Iciar 2008), Póvoa De Varzim, Portugal, 5112:527-536.

[23] Krishnamoorthi R (1998) A Unified Framework Orthogonal Polynomials For Edge Detection, Texture Analysis And Compression In Color Images, Ph.D. Thesis, 1998

[24] Krishnamoorthi R And Kannan N (2009) A New Integer Image Coding Technique Based On Orthogonal Polynomials, Image And Vision Computing, Vol 27(8). 999-1006.

Suguna R received M.Tech degree in CSE from IIT Madras, Chennai in 2004. She is currently pursuing the Ph.D. degree in Dept. of IT, MIT Campus, Anna University.

Anandhakumar P received Ph.D degree in CSE from Anna University, in 2006. He is working as Assistant Professor in Dept. of IT, MIT Campus, Anna University. His research area includes image processing and networks

A Comparative Study of Microarray Data Classification with Missing Values Imputation

Kairung Hengpraphrom[1], Sageemas Na Wichian[2] and Phayung Meesad[3]

[1]Department of Information Technology, Faculty of Information Technology

[2] Department of Social and Applied Science, College of Industrial Technology

[3] Department of Teacher Training in Electrical Engineering, Faculty of Technical Education

King Mongkut's University of Technology North Bangkok

1518 Piboolsongkram Rd.Bangsue, Bangkok 10800, Thailand

kairung2004@yahoo.com, sgm@kmutnb.ac.th, pym@kmutnb.ac.th

Abstract—**The incomplete data is an important problem in data mining. The consequent downstream analysis becomes less effective. Most algorithms for statistical data analysis need a complete set of data. Microarray data usually consists of a small number of samples with high dimensions but with a number of missing values. Many missing value imputation methods have been developed for microarray data, but only a few studies have investigated the relationship between missing value imputation method and classification accuracy. In this paper we carry out experiments with Colon Cancer dataset to evaluate the effectiveness of the four methods dealing with missing values imputations: the Row average method, KNN imputation, KNNFS imputation and Multiple Linear Regression imputation procedure. The considered classifier is the Support Vector Machine (SVM).**

Keywords;KNN, Regression, Microarray, Imputation, Missing Values

I. INTRODUCTION

Microarray data is a representative of thousands of genes at the same time. In with many types of experimental data, expression data obtained from microarray experiments are frequently peppered with missing values (MVs) that may occur for a variety of reasons, such as insufficient resolution, image corruption, dust, scratches on the slide, or errors in the process of experiments. Many data mining techniques have been proposed for analysis to identify regulatory patterns or similarities in expressions under similar conditions. For the analysis to be efficient, data mining techniques such as classification [1-3] and clustering [4-5] techniques require that the microarray data must be complete with no missing values [6]. One solution for the missing data problem is to go over the experiment again, but it is time consuming and very expensive [7]. Replacing the missing values by zero and average value can be helpful instead of eliminating the missing-value records [8], but the two simple methods are not very effective.

Consequently, many algorithms have been developed to accurately impute MVs in microarray experiments, for example K-Nearest Neighbor, Singular Value Decomposition, and Row average method have been proposed to estimate missing values in microarrays. KNN Impute was found to be the best among three methods [9]. However, there are still some points to improve. Many imputation techniques have been proposed to resolve the missing values problems. For example, Troyanskaya et al. [9] proposed KNN imputation based on Singular Value Decomposition and Row average methods. The results showed that KNN imputation method is better than the Row average method. Oba et al. [10] have proposed an imputation method called Bayesian Principal Component Analysis (BPCA). The researchers claimed that BPCA can estimate the missing values better than KNN and SVD. Another efficient method was proposed by Zhou et al. [11]. The method automatically selects gene parameters for estimation of missing values. The algorithm uses linear and nonlinear regression. The key benefit of the algorithm is quick estimation. Another research by Kim et al. [12] proposed local least squares (LLS) imputation. The idea is to use the similarity of structure of data as in least square optimization. This method is very robust. Later, Robust Least Squares Estimation with Principal Components (RLSP) was proposed by Yoon et al. [13] to improve the efficiency of the previous methods. RLSP imputation method showed better performance than KNN, LLS, and BPCA. The NRMSE is calculated to measure the imputation performance since the original values are now known.

Many missing value imputation methods have been developed for microarray data, but only a few studies have investigated the relationship between missing value imputation method and classification accuracy. In this paper, we carry out a model-based analysis to investigate how different properties of a dataset influence imputation and classification, and how imputation affects classification performance. We compare four imputation algorithms: the Row average method, KNN

imputation, KNNFS imputation and Multiple Linear Regression imputation method to measure how well the imputed dataset can preserve the discriminated power residing in the original dataset. The Support Vector Machine (SVM) is used as a classifier in this work.

The remainder of this paper is organized as follows. Section II provides theory and related works. The details of the proposed methodology are given in Section III. Section IV illustrates the simulation and comparison results. Finally, concluding remarks are given in section V.

II. RELATED WORK

A. Microarray Data

Every cell of living organisms contains a full set of chromosomes and identical genes. Only a portion of these genes are turned on and it is the subset that is expressed, conferring distinctive properties to each cell category.

There are two most important application forms for the DNA microarray technology: 1) identification of sequence (gene/gene mutation) and 2) determination of expression level (abundance) of genes of one sample or comparing gene transcription in two or more different kinds of cells. In data preparation, DNA Microarrays are small, solid supports onto which the sequences from thousands of different genes are attached at fixed locations. The supports themselves are usually glass microscope slides, the size of two side-by-side small fingers, but can also be silicon chips or nylon membranes. The DNA is printed, spotted, or actually synthesized directly onto the support. With the aid of a computer, the amount of mRNA bounding to the spots on the microarray is precisely measured, which generates a profile of gene expression in the cell. The generating process usually produces a lot of missing values and resulting in less efficiency of the downstream computational analysis [14].

B. K-nearest neighbor(KNN)

Due to its simplicity, K-Nearest Neighbor (KNN) method is one of the well-known methods to impute missing values in microarray data. The KNN method imputes missing values by selecting genes with expression values similar to the gene of interest. The steps of KNN imputation are as follows.

Step 1: Chose K genes that are most similar to the gene with the missing value (MV). In order to estimate the missing value x_{ij} of ith gene in j^{th} sample, K genes are selected whose expression vectors are similar to genetic expression of i in samples other than j.

Step 2: Measure the distance between two expression vectors x_i and x_j by using the Euclidian distance over the observed components in j^{th} sample. Euclidean distance between x_i and x_j can be calculated from (1)

$$d_{ij} = dist(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{1}$$

Where $dist(x_i, x_j)$ is the Euclidean distance between samples x_i and x_j; n is the number of features or dimensions of microarray; and x_{ik} is the kth feature of sample x_i.

Step 3: Estimate the missing value as an average of the K nearest neighbors, corresponding entries in the selected K expression vectors by using (2)

$$\hat{x}_{ij} = \frac{\sum_{k=1}^{K} X_k}{K} \tag{2}$$

$$X_k = X_{i=1...M} \mid d_i \in \{d_1, d_2, ..., d_K\}$$

where $\hat{x}_{ij}$ is the estimated missing value at i^{th} gene in j^{th} sample; d_i is the ith rank in distance of neighbor; X_k is the input matrix containing k^{th} rank in the nearest neighbor gene expressions; and M is the total number of samples in the training data.

C. The Algorithm of KNNFS

The algorithm of the combination of KNN-based feature selection and KNN-based imputation is as follows [15].

Phase 1: Feature Selection
    Step 1: Initialize K_F feature;
    Step 2: Calculate feature distance between X_j, $j = 1$, ..., *col* and X_{miss} (the feature with missing values) by using (1);
    Step 3: Sort feature distance in ascending order;
    Step 4: Select K_F minimum distances;
Phase 2: Imputation of Missing Values
    Step 5: Initialize K_C samples;
    Step 6: Use K_F feature to calculate sample distance between R_i, $i = 1$, ..., *row* and R_{miss} (the row with missing values) by using (1);
    Step 7: Sort sample distance ascending;
    Step 8: Select K_C minimum distance;
    Step 9: Use K_C sample to estimate missing value by an average of K_C most similar values by using (2).

D. Multiple Linear Regression

Multiple linear regression (MLR) is a method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is sometimes also called the predictand, and the independent variables are called the predictors.

The model expresses the value of a predictand variable as a linear function of one or more predictor variables and an error term:

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + ... + b_k x_{i,k} + e_i \tag{3}$$

$x_{i,k}$ is value of kth predictor in case i

b_0 is regression constant

b_i is coefficient on k^{th} the predictor

K is total number of predictors

y_i is predictand in case

e_i is error term

The model (3) is estimated by least squares, which yields parameter estimates such that the sum of squares of errors is minimized. The resulting prediction equation is

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} + ... + \hat{b}_k x_{i,k} \qquad (4)$$

Where the variables are defined as in (3) except that "^" denotes estimated values

III.   THE EXPERIMENTAL DESIGN

To compare the performance of the KNN, Row, Regression, and KNNFS imputation algorithms, NRMSE was used to measure the experimental results. The missing value estimation techniques were tested by randomly removing data values and then computing the estimation error. In the experiments, between 1% and 10% of the values were removed from the dataset randomly. Next, the four imputation algorithms as mention above are applied separately to calculate the missing values and then the imputed data (complete data) were used for accuracy measurement (NRMSE and classification accuracy) by SVM classifier. The overall process is shown in Fig. 1.
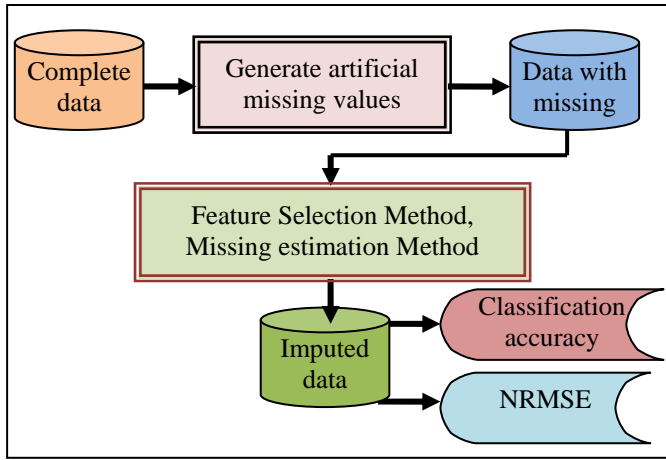
Figure 1.   Simulation flow chart.

To test the effectiveness of the different imputation algorithms, Conlon Cancer dataset was used. The data were collected from 62 patients: 40 tumor and 22 normal cases. The dataset has 2,000 selected genes. It is clean and contains no missing values.

The effectiveness of missing values imputation was computed by Normalized Room Mean Squared Error (NRMSE) [12] as shown in equation 5.

$$NRMSE = \frac{\sqrt{mean[(y_{guess} - y_{ans})^2]}}{std[y_{ans}]} \qquad (5)$$

Subject to

y_{guess} is estimated value

y_{ans} is prototype gene's value

$std[y_{ans}]$ is stand deviation of prototype gene

IV.   THE EXPERIMENTAL RESULTS

To evaluate the effectiveness of the imputation methods, the NRMSE values were computed using each algorithm as descript above. The experiment is repeated 10 times and reported the average as the result. The experimental results are shown in Tables I and Fig. 1.

Table I and Fig. 1 show the NRMSE of the estimation error for Colon Tumor data. The results show that the Regression method has a lower NRMSE compared to the other methods.

TABLE I.   NORMALIZE ROOT MEANS SQUARE ERROR OF MISSING-VALUE IMPUTATION FOR COLON CANCER DATA

% Miss	Colon Cancer			
	Row	*KNN*	*KNNFS*	*Regression*
1	0.6363	0.5486	0.4990	0.4049
2	0.6121	0.5366	0.4918	0.4103
3	0.6319	0.5606	0.5173	0.4282
4	0.6339	0.5621	0.5169	0.4251
5	0.6301	0.5673	0.5267	0.4410
6	0.6281	0.5634	0.5212	0.4573
7	0.6288	0.5680	0.5254	0.4415
8	0.6382	0.5882	0.5534	0.4548
9	0.6310	0.5858	0.5481	0.4418
10	0.6296	0.5849	0.5483	0.4450
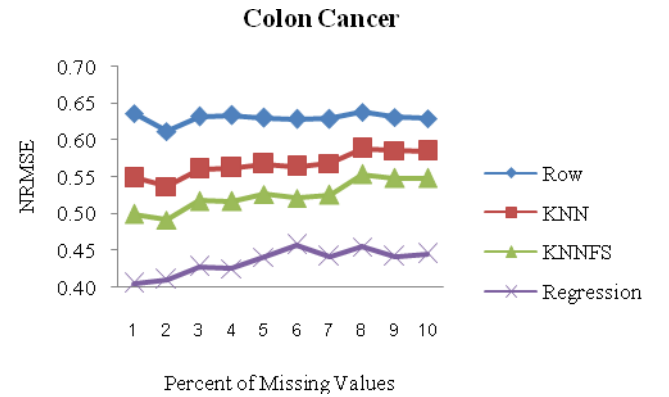
Figure 2.   Normalize root means square error of missing value imputation for Colon Cancer Data

The classification accuracy by using the SVM classifier is summarized in Table II and Fig. 2. The experimental results show that the accuracy of the row average method is ranged between 82.10% and 83.39%, while the neighbour-based methods (KNN, KNNFS) gave the result between 82.90% and 84.77%, and the regression method ranges between 82.90% and 84.84%.

V. CONCLUSION

This research studies the effectiveness of MVs imputation methods to the classification problems. The model-based approach is employed. Four methods for imputation (Row average, KNN, KNNFS, Regression) are used to compare the performance of classification accuracy in this research. The Colon Cancer data is used in this experiment.

To evaluate the performance of the imputation methods, we randomly removed known expression values between 1% and 10% of the values from the complete matrices, imputed MVs, and assessed the performance by using the NRMSE.

The results show that the Row average method yields a very poor effectiveness comparing with other methods in term of NRMSE. And also, it gives lowest classification accuracy with SVM classifier. For other methods, although the Regression yields the best performance in term of NRMSE, it is not different in classification accuracy.

TABLE II.   ACCURACY OF SVM CLASSIFYER FOR COLON CANCER DATA CLASSIFICATION

% Miss	Colon Cancer			
	Row	*KNN*	*KNNFS*	*Regression*
1	83.39	84.03	84.35	84.84
2	83.23	84.35	84.03	84.19
3	83.06	83.87	83.71	84.84
4	82.74	84.19	83.87	83.71
5	82.62	84.23	84.77	83.51
6	82.90	82.90	82.74	83.87
7	82.42	83.87	83.87	84.19
8	82.10	83.39	83.23	84.03
9	83.23	84.35	84.68	84.35
10	82.26	83.55	83.71	82.90
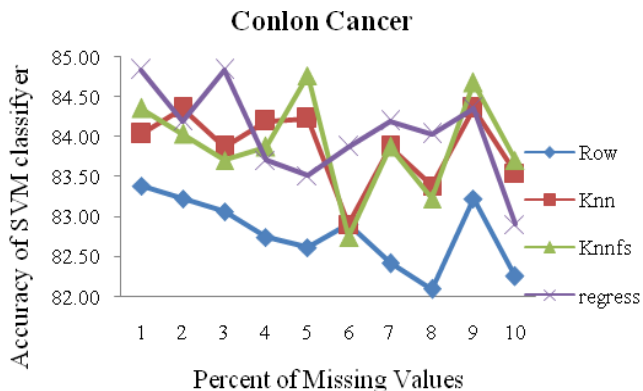
Figure 3.   Accuracy of SVM Classifyer for Colon Cancer

REFERENCES

[1]   M. P. S. Brown, W. N. Grundy, D. Lin , N Cristianini, C. W. Sugnet, T. S. Furey, M. J. Ares, D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines**",** Proc Natl Acad Sci USA, vol. 97, pp. 262-267, 2000.

[2]   X. L. Ji, J. L. Ling ,Z. R. Sun, "Mining gene expression data using a novel approach based on hidden Markov models", FEBS Letters, vol. 542, pp. 125-131, 2003.

[3]   O. Alter, P. O. Brown, D. Botstein, "Singular Value decomposition for genome-wide expression data processing and modeling", Proc Natl Acad Sci USA, vol. 97, pp. 10101-10106, 2000.

[4]   M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein , "Cluster analysis and display of genome-wide expression patterns", Proc Natl Acad Sci USA, vol. 97, pp. 262-267, 1998.

[5]   P. Tamayo, D. Slonim , J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky,E. S. Lander, T. R. Golub , "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation", Proc Natl Acad Sci USA, vol. 96, pp. 2907-2912, 1999.

[6]   E. Wit, and J. McClure, "Statistics for Microarrays: Design, Analysis and Inference", West Sussex: John Wiley and Sons Ltd, pp.65-69, 2004.

[7]   M. S. Sehgal, L. Gondal, L. S. Dooley, "Collateral Missing value imputation: a new robust missing value estimation algorithm for microarray data"**,** Bioinformatics, vol. 21, pp. 2417-2423, 2005.

[8]   A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt, et al.**,** "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling"**,** Nature, vol. 403, pp. 503-511, 2000.

[9]   O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R. B. Altman, "Missing values estimation methods for DNA microarrays", Bioinformatics, vol. 17, pp**.** 520-525, 2001.

[10] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. I. Matsubara, S. Ishii, "A Bayesian missing value estimation method for gene expression profile data", Bioinformatics, vol. 19, pp. 2088-2096, 2003.

[11] X. B. Xhou, X. D. Wang, E. R. Dougherty, "Missing –value estimation using linear and non-linear regression with Bayesian gene selection", Bioinformatics, vol. 19, pp. 2302-2307, 2003.

[12] H. Kim, G.H. Golub, H. Park**,** "Missing value estimation for DNA microarray gene expression data: local least squares imputation", Bioinformatics, vol. 21, pp. 187-198, 2005.

[13] D. Yoon, E. K. Lee, T. Park, "Robust imputation method for missing values in microarray data", BMC Bioinformatics, vol. 8, no. 2:S6, 2007.

[14] J. Quackenbush, "Microarray data normalization and transformation", Nature Genetics Supplement, vol. 32, pp. 496-501, 2002.

[15] P. Meesad and K. Hengpraprohm, "Combination of KNN-Based Feature Selection and KNNBased Missing-Value Imputation of Microarray Data", 2008 3rd International Conference on Innovative Computing Information and Control, pp.341, 2008.

Dependability Analysis on Web Service Security: Business Logic Driven Approach

Saleem Basha

Department of Computer Science

Pondicherry University

Puducherry, India

smartsaleem1979@gmail.com

P. Dhavachelvan

Department of Computer Science

Pondicherry University

Puducherry, India

dhavachelvan@gmail.com

Abstract— **In the modern computing world internet and e-business are the composite blend of web service and technology. Organization must secure their state of computing system or risk to malicious attacks. The business logic is the fundamental drive for computer based business tasks, where business process and business function adds their features for better illustration for the abstract view of the business domain. The advent and astronomical raise of internet and ebusiness makes the business logic to specify and drive the web service. Due to the loosely coupling of web service with the application, analyzing dependability of the business logic becomes an essential artifact to produce complex web service composition and orchestrations to complete a business task. This paper extended the Markov chain for the dependability analysis of the business logic driven web service security.**

Keywords- Web Servcie; Dependability Analysis; Busienss Logic; Web Servcie Security

I. INTRODUCTION

Enterprise systems are distinct and highly complex class of systems. They are characterized by their importance for enterprises themselves, making them mission critical, by their extreme multi-user capability, by their tolerance of heavy loads and by with their tight integration with the business process, which makes every enterprise system installation unique. In short, they are one of the most fascinating yet most demanding disciplines in software engineering [1]. The business logic is responsible for implementing the basic rules of the system according to the operating rules of the business. Its main feature is to take request, determine what actions the request requires, implement those actions and return response data to the customer. Organization faces the problem of the security derived from the non functional requirements and to maximize the utilization of the cutting edge technology with minimum cost in the agile business environment. Web service is the upcoming wave for tomorrows business needs, in this concern the non functional attributes is the one of the major challenging sector for the developers to guarantee the confidentiality, authentication, integrity, authorization and non-repudiation of machine to machine interaction so security is not negotiable to anticipate a secure artifacts for web service. There are two underlying themes for all these pressure: Heterogeneity and agility: Software development is a standard practice in software engineering where business logic drives the software

development starting from requirement analysis to maintenance. The information exchange between the database and the user interface will be done by the functional algorithm which is described by the business logic. This logic is composed of business functions and business rules. Series of logically related activities or task performed together to produce a defined set of result called business function and business rule is a statement that defines or constrains some aspect of the business. It is important to understand that business modeling commonly refers to business process design at the operational level [4] which comes under the functional requirement of the system, where as the non functional requirements are left as it is afterthought. Non functional attributes defines the system properties and constraints and can be classified as Product requirements, Organizational requirements and External requirements. Security of the system plays a major role across the boundaries of the organizations. Security of the system can be improved by providing the foundation in the early phase of the system development process by dependability analysis. The development of system during requirements analysis and system design can improve the quality of the resulting system.

The most common dependability parameters which can be used to describe the nonfunctional requirements of virtually any kind of service, independently from the nature of the service are reliability and availability [20]. The dependability of the of the system raises along with the growing popularity of the web service based integration of heterogeneous enterprise systems. The parameters of non functional (mainly dependability related) requirements must be predefined for a given web service in order to guarantee the web service consumers. The provider also has to consider similar nonfunctional parameters of external Web services involved in the operation of his main service to be able to calculate and plan the dependability parameters.

In this paper, we extend Markov chain process for the dependability analysis of the business logic driven web service security. A direct generalization of the scheme of independent trials is a scheme of what are known as Markov Chains, imagine that a sequence of trials in each of which one and only one of k mutually exclusive events $A_1^{(s)}$, $A_2^{(s)}$… $A_k^{(s)}$ can occur. We say that the sequence of trials forms a Markov Chain, or more precisely a simple Markov chain, if the conditional

probability that event $A_i^{(s+1)}$ (i=1,2…k) will occur in the $(s+1)^{th}$ trial (s=1,2,3….) after a known event has occurred in the sth trial, depends solely on the event that occurred in the sth trial and is not modified by supplementary information about the event that occurred in earlier trials. A different terminology is frequently employed in starting the theory of Markov chains and one speaks of a certain system S, which at each instant of time can be in one of the states A_1, A_2 ….. A_k and alters its state only at times t_1, t_2 …. t_n …. For Markov chains, the probability of passing to some state A_i (i=1,2…k) at time $\tau(t_s < \tau < t_s+1)$ depends only on the state the system was in at time $t(t_{s-1} < t < t_s)$ and does not change if we learn its state were at earlier times.

II. WEB SERVICE SECURITY ANALYSIS AND BUSINESS LOGIC MODEL

Modeling business logic focuses on the core functionality of the business process, which are capsulated as web services. It requires that business process pertains exactly to the business logic with various business terminologies such as dependency, policy, standards, constraints, etc. As a prerequisite to this business logic model, the core functionality of the business process should be analyzed for dependencies then modeled absolutely, whereas the previous implementations of web services were direct. Ronald et al. states that existing models like business rule model, business motivation model and business process model concentrate on business process at the operational level with compromising minimum range of QoS attributes [2]. Business rule model deals with the extraction of business rules from the business logic, in order to reduce the cost and time spent in development [2][3]. Business motivation model paves way for identifying the facts preserved in novel objectives, thereby facilitating the business process development. Business process model provides optimization to the business process at the designing phase. The implementation of a company's business model into organizational structures and systems is part of a company's business operations. It is important to understand that business modeling commonly refers to business process design at the operational level [4], whereas business models and business model design refer to defining the business logic of a company at the strategic level. Business logic model aims to resolve the complexities involved, by decomposing the business process into sub processes and in turn into tasks, also preserving the functional dependencies among the sub-processes, without ignoring the key factors. Any service domain adopted this model for their web service development could be easily managed in terms of handling run time exceptions towards service reliability and manageability. Business logic model can be applied in tandem with the above described models, thereby facilitating service computation and composition much better. This model enables web services to realize their computational criteria such as computability, traceability and decidability with the supporting QoS attributes like manageability, configurability, serviceability and dependency. The computational criteria would be the best suit for the web service community who look for exception-free web services or reconfigurable web services. This model would also satisfy the service consumers who approach the discovery and composition engines for fetching exception free or self

configurable web services. Hence this model would ensure the consumers that the services are manageable at runtime, self configurable in case of dependability, computable in total or partial and traceable to the point of failure. Also it sustains dependency between the business rules and business functions.

A. Web Service Security Analysis

The cost versus risk parameters of the business will determine the capability to implement security in web service [25]. More a business can articulate the risks to its business, better it will be capable to appraise the advantage of preventive measurements to protect itself. The business must be capable of answering such a question.

Who has to have the access to which information?

How is access to data provided? Direct or brokered?

Is there a need for data to be available to external partners as well as internal consumers?

What requirements does the information need in transit, in process and at rest?

To achieve a secure web service, the application and the security analysis must be analyzed conceptually and modeled. This roughly goes without saying that the big companies are obsessed by the safety and to assure the critical applications, essential information is at stake. Any movement towards web service presents a principal opportunity to incorporate the safety in future applications. Organization and system stake holders are realizing that every opportunity for the business emerges with the danger of seriously screwing things-up. In early web service adopters are delicious prey for the bad thinking about the security analysis of the web service. After the several advancement in the technology and techniques in the context of security analysis, still the system developers faces the problem of security and security analysis.

Wide consideration to inherent the security features in the SDLC of the web service platform will enhances the safety of the web service as well as the service themselves [26]. Thus web service provides an opportunity to avoid such security related issues and challenges or otherwise managing security dependencies that pervade software architecture.

The vendors typically emphasize the primary features of safety that they offer as key selling points in the real world of enterprise applications. Nevertheless, out of the list of obligatory features of safety, few sellers can give testimony to the underlying safety of the product itself. So the user could have all the characteristics of security in the computing world, but they remain untenably insecure due to lack of analysis of the security.

B. Business Logic Model

Business processes and motivation models have been used to analyze and propose new changes in accordance to changing business scenarios. A process model scope does not extend optimally to web services, whereas Business Rule models extract rules from the business logic and concentrate mainly on the problem of modeling and accessing data by using efficient queries [4][2]. However they do not model the entire business

logic including the dependency analysis. Thus there is a need for a model which represents a business process in detail and also adapts the dependability analysis, rules, policies and standards to changing business scenarios. This adaptability helps service consumers and service providers cope up with the demanding and challenging changes in services.

Such a representation should not compromise on matter and processes private to a business. Since a business logic model seems inevitable, by maintaining business privacy and by modeling a specific business process, the model seems to be a promising methodology to handle the ever-changing business scenarios. Business Process systems that use web services decrease the cost of automating transactions with trading partners.

The scope of a business process is limited to design, development and deployment of services. The limited scope helps to develop better services keeping service customization in mind. The outcome breakdown structure of the service business logic is streamed as a set of business rules, functions and parameters. Further, these rules and functions could be tuned to be primitive business functions under certain specific conditions. The primary motivation behind setting up the business functions as primitive business functions would pose the computability and traceability factors, which are the most essential quality-driven factors as they could manage the complete service computing platform successfully by the effective handling of run-time exceptions during service computation and composition by the security dependencies. This model decomposes the business logic into functionally consistent and coherent business rules and functions, keeping in mind the privacy constraints of businesses. Decomposition helps representing the interdependent business functions with the security dependability as low as possible. This strategy categorizes the business functions into initial, composite and recursive functions and evaluates them into computable and dependable business functions. Computability and dependability of business functions are key factors for measuring the success rate. Existing discovery and composition engines provide services based on functionality, quality, and security of requested services. Customizing the services is not addressed by the existing engines. The proposed business logic based dependability analysis exhibits the functionalities of any of the generic engines but is also resilient to customization.

C. Relation Between Web Service Security Analysis and Business Logic Model

Modeling system with business logic model has benefits like; it reflects standard layering practices with in the development communities, business functionality easily accessible by other object application, very efficient to build business objects, it helps to test the basic success premises of business, improves the clear understanding of existing value drivers and constraints, it provides a componentized view of the business and technology environment in order to have common building blocks that can be reused across product and business silos, it defines and sustainable interim states which provides measurable benefits as flexible path to the goal and business logic provides a strong governance to manage and deliver the changes. Business logic also possesses some of the

drawbacks; significant performance problem for data intensive functions, non object application may have significant difficulty to accessing functionality. Improper handling of the non functional requirements and its dependability may result in compromising the growth of the organization.

Currently much work in the requirements engineering field has been done to shown the necessity of business logic which take non-functional requirement's (NFR) dependability into consideration. Such logic will better deal with real-world situations. On the other hand the advantages of having business logic is the capability of representing nonfunctional aspects, such as dependability, confidentiality, performance, ease of use and timeliness. It is believed that these functional aspects should be dealt with as non-functional requirements. Therefore, NFRs have to be handled and expressed very early in the process of modeling an information system [5]. Organizations are spending much in system development and least concentration to NFRs. Recent tales of failure in information systems can be explained by the lack of attention to NFRs. The London Ambulance System (LAS) is a example for the information system failure due to lack of attention of NFRs [6]. The LAS was deactivated, soon after its deployment, because of several problems, many of which were related to NFRs such as performance and conformance with standards [7]. Negotiation in the NRFs is not a healthy activity in the system development, the consequences of negotiating NRFs leads to serious problem as in the case of LAS.

Serviced Oriented Architecture (SOA) is the paradigm for the future business environment, where web service is the building block for SOA and it is the key for agile business across the enterprises. It is important in Service Oriented Architecture to separate functional and non-functional requirements for services because different applications use services in different non-functional contexts. In order to maximize the reusability of services, a set of constraints among non-functional requirements tend to be complicated to maintain. Currently, those non-functional constraints are informally specified in natural languages, and developers need to ensure that their applications satisfy the constraints in manual and ad-hoc manners [8]. System developers believe that business logic composes and speaks only the functional aspect, but fails to keep in mind that to consider the other aspects driven by functional aspect i.e. dependabelity. The separation of functional and non-functional aspects improves the reusability of services and connections. It also improves the ease of understanding application design and enables two different aspects to evolve independently. Wada et al. pointed that the separation of functional and non-functional aspects results in higher maintainability of applications [9]. Non-functional aspects should also be captured as abstract models in an early development phase and automatically transformed to code or configuration files in order to improve development productivity. It incurs time-consuming and error-prone manual efforts to implement and deploy non-functional aspects in later development phases (e.g., integration and test phases) [10][11]. Web services become more popular and better utilized by many users and software agents, they will inevitably be commercialized. But still Services Challenge (WSC) that focus on functional aspects [12][13]. We believe that considering the

dependability of both functional and non-functional attributes together in solving the Web services composition problem would produce superior outputs [14]. Because NFRs are always tied up with functional requirements i.e., NFRs can be seen as requirements that constrain or set some quality attributes upon a functional requirement[25]

To the best of our knowledge, this is the first work studying the usability of the main approaches adopted for specifying and enforcing web service security analysis in business logic. Today's internet and e-affaires are the composite blend of business process and technology where the web service is the perfect blue print for agile business environment. In the early times, data in the networks were closed; security within these networks was ensured through isolation. Later LAN(Local Area Network) was introduced with firewalls to isolated from the untrusted public networks to ensure that adversaries and hackers cannot intrude into the private network. For more security, they added security aspects like proxies, intrusion detection system, intrusion prevention system, antivirus, malware catchers etc., are the domain specific security measures. The belief was that applications and assets used by the organization can be secured through in-vitro perimeter security. Therefore, software engineering techniques never looked into security analysis as an important component in Software Develop Life Cycle (SDLC); and, identified security as nonfunctional requirement [15]. Security must be part of the application to protect itself from security threats. Application security will however be over and above the perimeter network security. To achieve this, security now need to be treated as functional requirement and must be part of SDLC [16]. Sindre et al. have identified application security as a need and proposed ways to achieve this. All these isolated and independent techniques have been combined together in a thread to form a business Logic [17].

III. DEPENDABILITY ANALYSIS IN WEB SERVICE SECURITY

A. Dependability Analysis

The most common dependability parameters which can be used to describe the nonfunctional requirements of virtually any kind of service, independently from the nature of the service are reliability and availability [20]. The probability formalism, into which these dependencies may fit in a natural way and it is important for the analysis of the non functional parameters. Then the dependability of the system can able to assessed for the parameters of the system from the components' parameters. Using design patterns that are proven in the field of reliability can enhance the dependability of the main service. Such patterns can be, for instance, the N-Version Programming and the Recovery Block scheme [18]. Web service is the building block for SOA in different platforms, vendors, etc. The dependability of that particular system may of course influenced by the nature of the problem. The parameters of a composite web service is depends on the nature of the implementation and design of the individual web services and its patterns. Finally the aim of the dependability analysis of the system is to validate a business process towards some business tasks. The consideration of such patterns can be based on the result of a dependability analysis, moreover the

constrains of the system can be eliminated and the probability of a system failure can be evaluated. The analysis can be done for two basic purposes; determine the optimal solution for given requirements and determine the guaranteed parameters for a given solution.

B. Business Logic Based Dependability Analysis in Web Servcie Security

The web service is the perfect blue print for agile business environment where the services are catered across the organizational boundary which is specified by the business logic. The loosely coupling characteristic of web service introduces many challenges including security.

Security is the major concern and web service may fail due to these concerns. As said earlier business logic drives the business through web service using business functions and business rules. Business logic also specifies the security aspects; a promising approach for problem determination in large systems is dependency analysis. In brief, the question that dependency analysis tries to answer is this: Is the service X dependent on another service Y or security parameter Z? If such a dependency exists, what is the strength1 of the dependency? Using this information, when a problem is observed at a particular service point, the root cause may be tracked down to a security parameter on which this service is dependent. The dependency analysis problem becomes very challenging in situations where the security of the system may be static or dynamic in nature. In such cases, these parameters can appear and disappear during system lifetime because of failures, or deployment of new security requirement and the dependency relations can change as a result of change of security parameter availability or new service level agreements being negotiated.

For illustration let us consider four service providers (SP1, SP2, SP3, SP4) each service provider has his own Business Logic (BL) and one or many Business Function (BF) to complete the business tasks as shown in the Figure 1.

From the Markov chain the dependability of the business functions to the web service is shown in the Table 1. The BL_1 has defined two business functions namely BF_1 and BF_2 which has three web services each WS_1, WS_3, WS_4 and WS_1, WS_2, WS_4 respectively. Now consider only the business function BF_1, let WS_1, WS_3 and WS_4 are need to complete a business task with some security consideration. The state graph of these web services is show in the Figure 2. WS_1 is the initial state or the initial web service for BF_1, the arrow flows from WS_1 to WS_2 iff (if and if only) all the security conditions satisfies in WS_1, and its probability is 1, else it rolls back to WS_1 itself. Similarly from WS_2 to WS_4, the P_{21} is the probability of the state WS_3 to return to previous state WS_1 under any fault conditions, and P_{23} is the probability of success of the security considerations and reaches to the final state WS_4 and thus a business task completes for business function BF_1.

(a)

(b)

(c)

(d)

Figure 1.   Service Providers (SP$_1$, SP$_2$,SP$_3$ and SP$_4$) and its Business Functions

TABLE I.   BUSINESS LOGIC AND ITS ASSOCIATED BUSINESS FUNCTIONS AND WEB SERVCIES

	BL$_1$		BL$_2$			BL$_3$			BL$_4$
	BF$_1$	BF$_2$	BF$_3$	BF$_4$	BF$_5$	BF$_6$	BF$_7$	BF$_8$	BF$_9$
WS$_1$	*	*							
WS$_2$		*							
WS$_3$	*								
WS$_4$	*	*							
WS$_5$			*		*				
WS$_6$			*	*					
WS$_7$				*	*				
WS$_8$						*	*	*	
WS$_9$							*		
WS$_{10}$						*	*		
WS$_{11}$									*
WS$_{12}$									*

● ● ●   are the security considerations

Figure 2.   Dependability Graph of BF$_1$
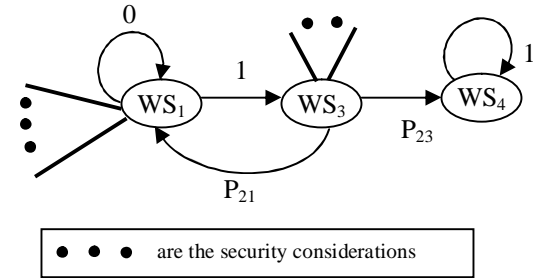
The transition probability of BF1 from state WSi to WSj, where i, j = 1,3,4. Then transition matrix can be written for WS1, WS3, and WS4.

$$BL_1 = \begin{pmatrix} 0 & 1 & 0 \\ P_{21} & 0 & P_{23} \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

Here P$_{21}$ + P$_{23}$ = 1

Let $\partial_0, \partial_1, \partial_2 \ldots.. \partial_n$ are the phases of the chain, then

P$_i$ = [P$_1^{(i)}$ P$_2^{(i)}$ P$_3^{(i)}$] be the probability of the chain in the given phase i.

Since WS$_1$ is the initial state, therefore P$_0$ = [1 0 0]

Further from matrix theory P$_{i+1}$ = P$_i$A i.e.

P$_1$ = P$_0$A = [010]

P$_2$ = P$_1$A = [P$_{21}$0P$_{23}$]

P$_3$ = P$_2$A = [0P$_{21}$+P$_{32}$P$_{23}$1]

In general P$_n$=P$_0$A^n ; n=1,2,3 ….

where

$$A = \begin{pmatrix} 0 & 0 & 1 \\ P_{23} & P_{21} & 0 \\ 1 & 0 & 0 \end{pmatrix} \qquad (2)$$

$$A^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ P_{23} & P_{21} & 0 \end{pmatrix} \qquad (3)$$

The matrix of non absorption states is represented as Q

$$Q = \begin{pmatrix} 0 & 1 \\ P_{21} & 0 \end{pmatrix} \qquad (4)$$

From the matrix theory $I_3 - Q$ is always invertible matrix which is the fundamental matrix N of the chain is given by

$$N = (I_3 - Q)^{-1} = \frac{1}{D(I_3 - Q)adj(I_3 - Q)} \qquad (5)$$

Then ij^{th} entry of the N gives the mean time of that state. For example, assume that there are four business functions which is provided by a service provider in association with three web services (WS$_1$, WS$_2$, and WS$_3$), first business function (BF$_1$) has 6 dependencies, second business function (BF$_2$) has 54 dependencies, third business function (BF$_3$) has 28 dependencies and fourth business function (BF$_4$) has 9 dependencies over those web services to complete a business task with 4 phases of Markov chain. Then the state transition matrix of these web services can be given as for the completion of a business task with minimum dependencies is given below. Assume that the business logic with respect to the particular web service to fulfill a business task could be produced statistically is shown in the matrix below.

$$Bu\sin essTask = \begin{matrix} & BF_1 & BF_2 & BF_3 & BF_4 \\ WS_1 \\ WS_2 \\ WS_3 \end{matrix} \begin{pmatrix} 3 & 8 & 5 & 4 \\ 4 & 2 & 6 & 5 \\ 2 & 1 & 1 & -1 \end{pmatrix} \quad (6)$$

- The dependencies of BF$_1$ over WS$_1$ is 3, WS$_2$ is 4 WS$_3$ is 2.

- The dependencies of BF$_2$ over WS$_1$ is 8, WS$_2$ is 2 WS$_3$ is 1.

- The dependencies of BF$_3$ over WS$_1$ is 5, WS$_2$ is 6 WS$_3$ is 1.

- The dependencies of BF$_1$ over WS$_1$ is 4, WS$_2$ is 5 WS$_3$ is -1.

- The row total of WS$_1$ for the four business functions are 20.

- The row total of WS$_2$ for the four business functions are 17.

- The row total of WS$_3$ for the four business functions are 3.

From the above matrix it is clear that BF$_4$ has the minimum dependencies that the other three business tasks. The state transition diagram of the business task is given as states in the Figure 3.
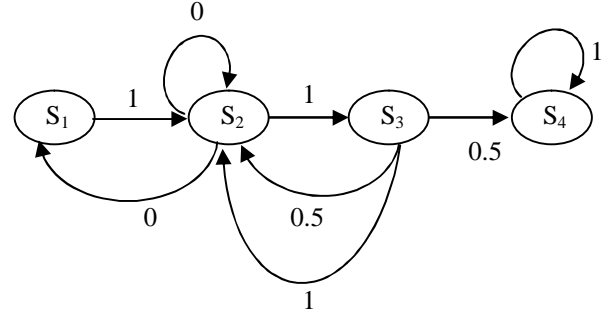
Figure 3. State representation of BF$_1$, BF$_2$, BF$_3$ and BF$_4$

Considering the other three business logics, P21=0, P23=1, P34=0.5 and P32=P34=0.5.

$$N = \frac{1}{0.5} \begin{pmatrix} 0.5 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0.5 & 1 \end{pmatrix} \qquad (7)$$

Therefore the total dependencies are 1*6+2*54+2*28+1*9=179 for 5 phases. For 4 phases it is given as P3=[0 0.5 0 0.5] ; P4(3) = 0.5. Hence to complete a business task in four phases it has only the probability of 50%.

The starting chain is S$_i$, then the expected number of steps before the chain is absorbed is given by, let ti be the excepted umber of steps before the chain is absorbed, t be the column vector whose ith entry is t$_i$.

$$t = Nc \qquad (8)$$

where, c is a column vector all of whose entries are 1

$$t = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 2 & 2 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix} \qquad (9)$$

1) Classification of Possible States

In a Markov chain, each state can be placed in one of the three classifications. Since each state falls into one and only one category, these categories partition the states. The secret of categorizing the states is to find the communicating classes. The states of a Markov chain can be partitioned into these communicating classes. Two states communicate if and only if it is possible to go from each to the other. That is, states A and

B communicate if and only if it is possible to go from A to B and from B to A. There are three classification of states transient, ergodic, and periodic.

The state A_i is called transient if there exist A_j and n such that $P_{ij}(n)>0$, but $P_{ij}(m)=0$ for all m. Thus, an transient state possess the property that it is possible, with positive probability, to pass from it into other state, but it is no longer possible to return from that state to the original state.
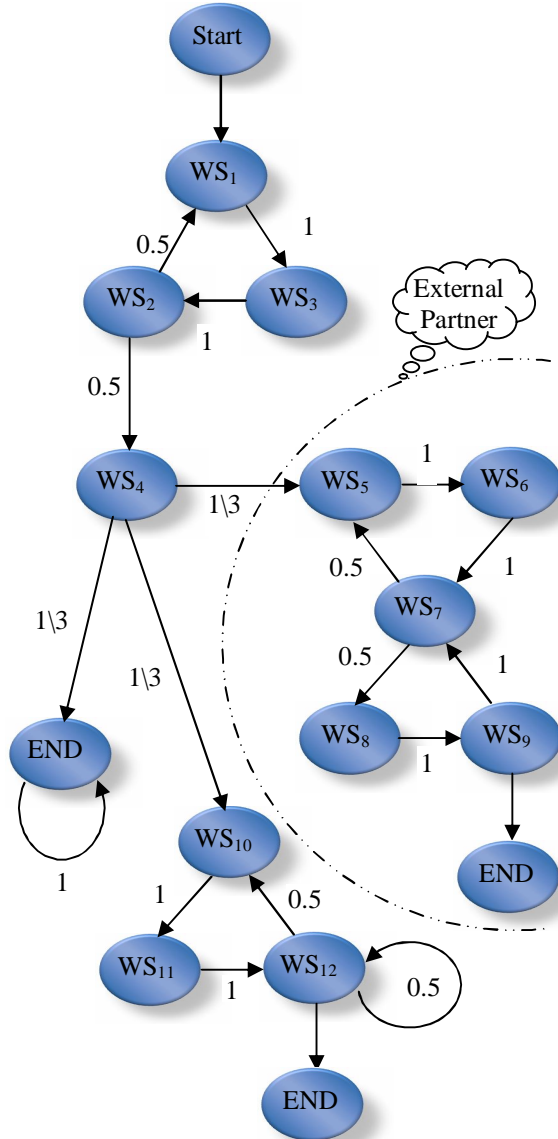
Figure 4.   Sample classification of concerns

All states not transient are called periodic state. Form the definition it follows that if the states A_i and A_j are essential, then there exist positive m and n such that as long with the inequality $P_{ij}(m)>0$ the inequality $P_{ij}(n)>0$ also holds. If A_i and A_j are such that for both of them these inequalities holds, given certain m and n, then they are called communicating. It is clear that if A_i communicates with A_j, and A_j communicates with A_k, then A_i also communicates with A_k. Thus, all essential states can be partitioned into classes such that all states belonging to a single class communicates and those belonging to different class do not communicate. Since for the essential state A_i and the unessential state A_j the equation $P_{ij}(m)=0$ holds for any m, we can draw the following conclusion: if a system has reached one of the states of a definite class of essential states, it can no longer leave that class.

Transient: A state is transient if it is possible to leave the state and never return.

Periodic: A state is periodic if it is not transient, and if that state is returned to only on multiples of some positive integer greater than 1. This integer is known as the period of the state.

Ergodic: A state is ergodic if it is neither transient nor periodic.

The Figure 4 illustrates the classification of the states for a banking transaction. For illustration assume there are two service providers SP_1 and SP_2. SP_1 has the set of web services (WS_1, WS_2, WS_3, WS_4, WS_{10}, WS_{11}, WS_{12} and Event Notification EN) and the SP_2 has another set of web services (WS_5, WS_6, WS_7, WS_8 and WS_9) which is under the dotted circle, the web services can be noted as states of the transactions. WS_1, WS_2 and WS_3 are the basic transactions which are communicating class. Neglecting start and end, once the chain goes from WS_1 to WS_4 it cannot return to WS_1, hence the web services WS_1, WS_2, WS_3 are transient. WS_4 acts as a gateway for the external partners. Web service WS_4 is a communicating class by itself, once the control leaves WS_4 it never returns again to WS_4 so the web service WS_4 is transient. Any failure occurs in the gateway will be captured by the EN and notified as an event notification. The EN is a communication class and has the loop so it is ergodic. WS_{10}, WS_{11} and WS_{12} be the loan approval services, WS_{12} is the final web service which decide the approval process base upon the parameters passed by the other web services and finally ends the process else it rollbacks. The web services WS_{10}, WS_{11} and WS_{12} forms a communicating class. Once the control arrives there it never leaves the class so it is not transient, also the web service WS_{12} has a loop it and its whole class cannot be periodic hence it is ergodic.

The external partner has five web services which forms a communicating class. Once the control comes to this class it never leave that class hence they are not transient if we consider the web service WS_7 once the control leaves WS_7, will always return in 3 transitions hence the whole class forms a periodic.

Let us examine more closely the mechanism of transition from state to state inside on class. To do this take some essential state A_i and denote by M_i the set of all web services WS for which $P_{ii}(WS)>0$. This set cannot be empty by the virtue of the definition of an essential state. It is immediately obvious that if the web service WS_i and WS_j are contained in the set M_i, then their dependability, of WS_i and WS_j, also belongs to this set. Denoted by d_i the greatest common dependability of the entire web services of the set M_i. it is clear that M_i consists only of web services which are dependents of d_i. The dependencies d_i is called the period of the state A_i.

2) Limiting probabilities of composite web service

In a service-oriented architecture [21], individual services are combined into a single workflow that reflects the business process in question. Although services can be defined in a general way, in practice the most widely used services are web services [22][23]. Currently, composition of web services is carried out by orchestration [24]. An orchestration is a workflow that combines invocations of individual operations of the web services involved. It is therefore a composition of individual operations, rather than a composition of entire web services.

The greatest probabilities $P_{ij}(n)$ cannot increase with the growth of n and the least cannot decrease, where n is the composite factor (no. of web services to form a composite web service) in other words, the group of communicating web services in a class is called composite web service. It is then shown that the maximum of the difference $P_{ij}(n) - P_{lj}(n)$, (i,l = 1,2,3....k) tends to zero when n tends to infinity. It is cleared that the when the number of web services (composite factor) increases in the composite web service, then the probability of change of state decreases to zero. Then there exist

$$\lim_{n\to\infty}.\min_{1\le i\le k}.P_{ij}(n) = \overline{P_j} \qquad (10)$$

and

$$\lim_{n\to\infty}.\max_{1\le i\le k}.P_{ij}(n) = \overline{\overline{P_j}} \qquad (11)$$
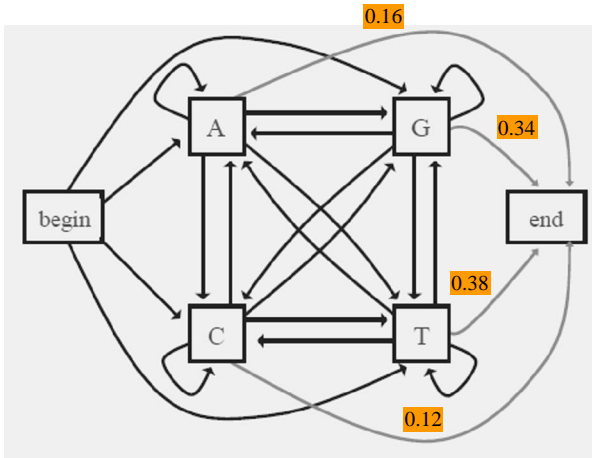
Figure 5.   Composite web service

From the Figure 5, let A, C, G and T be the web individual web service to form a composite web service and they are communicating class with the composite factor 4. Each individual web service has its own security constrains and it is marked as self loop. Start state is the initial orchestration of web service to do a business task and end state is the final work done by the orchestration.

Then the transition probabilities

$P_r(X_i=end \mid X_{i-1}=A) = 0.16$

$P_r(X_i=end \mid X_{i-1}=G) = 0.34$

$P_r(X_i=end \mid X_{i-1}=T) = 0.38$

$P_r(X_i=end \mid X_{i-1}=C) = 0.12$

$$P_r = \begin{pmatrix} 0.16 & 0.34 \\ 0.38 & 0.12 \end{pmatrix} = 0.049096 \qquad (12)$$

If the composite factor is reduced by 2 then the transition probabilities

$P_r(X_i=end \mid X_{i-1}=G) = 0.34$

$P_r(X_i=end \mid X_{i-1}=A) = 0.16$

$$P_r = \begin{pmatrix} 0.34 & 0.16 \\ 0 & 0 \end{pmatrix} = 0.18 \qquad (13)$$

Therefore, The probability of changing state from start to end in a composite web service with the composition factor 4 is 0.049096 and the probability of changing state from start to end in a composite web service with the composition factor 2 is only 0.18. Hence it is concluded that the probability to complete a business task for a given composite web service inversely proportional to the number of individual web service (composite factor).

Defining the composite service with very small composite factor will increase the probability to complete the business task and also supports reusability & flexible-introduces governance, maintenance & new testing, performance issue based on the network consumption of these service.

Defining the composite service with too large composite factor will decrease the probability to complete the business task and also deliver less or no reusability & flexibility but easy to maintain with less network usage.

Finding the right choice of composite factor is on of the key success factor to web service computing

IV.   CASE STUDY / MODEL ANALYSIS

Dependability analysis is unavoidable in service computing and hence, analyzing these expendabilities could resolve these problems up to the maximum extent. The purpose of analyzing these dependencies is to ensure that the code can handle any exception or error during the service is being computed. The service computation in this context is also about when more number of services is executed under service composition. Table II illustrates the real world web service and its dependencies.

TABLE II.   DEPENDABILITY OF WEB SERVICE SECURITY

Web Service Endpoint	Service Functionality	Business Logic Dependabilities
http://xml.assessment.com/service/MAPPMatching.asmx?wsdl	Match a single Job Profile to a single person	1. Multi criteria and profile match doesn't set to the service 2. No multi value dependency Exist
http://www.strikeiron.com/webservices/usdadata.asmx?wsdl	StrikeIron provides an ondemand Web-based infrastructure for delivering business data to any Internetconnected system.	1.Requested type of data delivery is not applicable 2. Data source not found 3.Null pointer exception
http://www.holidaywebservice.com/Holidays/GBNIR/Dates/GBNIRHolidayDates.asmx?WSDL	Web service that calculates specific national holidays for Northern Ireland (UK)	1.Invalid date format 2. No match exist
http://galex.stsci.edu/casjobs/CasUsers.asmx?WSDL	Login web service uses either name or email id	1.Null pointer Exception 2. Can't resolve the input Symbol
http://websrv.cs.fsu.edu/~engelen/interop2_2.wsdl	Service for typecasting includes hexadecimal, base64,etc	1. implicit type conversion from type1 to type2 not Possible
http://trial.serviceobjects.com/pt/PackTrack.asmx?wsdl	Package tracking service : Input all digits of the package tracking number. Returns package tracking information for a given Airborne Express number	1. Data Mismatched found 2. Duplicate package number exist 3. Data inconsistency
http://superglue.badc.rl.ac.uk/exist/services/Discovery?wsdl	Provides simple and fast information retrieval for the given input string.	1. unhandled exception 2. resource not found

V.   CONCLUSION

The exploit of web threats continues to expand and security concerns wane in their usefulness. The current workflow modeling and integration software are not able to capture important non-functional parameters of the system, like security dependability which is crucial with the model transformation framework. Probability analysis of the security dependencies represents another step in this direction such as Markov chain. In this paper we extended the concept of Markov chain process for dependability analysis of business logic for web services. The presented approach is fully base on mathematical concepts and modeling of business logic dependability analysis of web service security can be seamlessly integrated to business logic analyzing algorithms.

ACKNOWLEDGMENT

REFERENCES

[1] Dirk Draheim, Gerald Weber, "From-Oriented Analysis, A New Methodology to Model Based Application", Springer, vol 4(3), pp 346-347, 2005

[2] Ronald G.Ross, "Principles of the Business Rule Approach", Addison Wesley Publisher, ISBN 0-201-78893-4, 2003.

[3] Asuman Dogac, Yildiray Kabak, Tuncay Namli, and Alper Okcan, "Collaborative Business Process Support in eHealth: Integrating IHE Profiles Through ebXML Business Process Specification Language", IEEE Transactions on Information Technology in Biomedicine, vol. 12(6), pp 754-762, 2008.

[4] Saqib Ali, Ben Soh, and Torab Torabi ,"A Novel Approach Toward Integration of Rules Into Business Processes Using An Agent-Oriented Framework" , IEEE Transaction on Industrial Informatics, Vol. 2(3), pp 145-154, 2006.

[5] Luiz Marcio Cysneiros, Julio Cesar Sampaio do Prado Leite and Jaime de Melo Sabat Neto, "A Framework for Integrating Non-Functional Requirements into Conceptual Models" Springer. LNCS, Issue 2068, pp 284-298, 2001.

[6] Finkelstein A, Dowell J, "A Comedy of Errors: the London Ambulance Service Case Study" Proceedings of the Eighth International Workshop on Software Specification and Design, IEEE Computer Society Press, pp 2-5, 1996.

[7] Breitman KK, Leite JCSP, Finkelstein A. "The world's Stage: A Survey on Requirements Engineering Using a Real-Life Case Study" Brazilian Computer Society, pp 13-37, 1999

[8] Wada. H, Suzuki. J and Oba. K "A Feature Modeling Support for Non-Functional Constraints in Service Oriented Architecture" IEEE Conference on Service Computing, pp 187-195, 2007

[9] Wada. H, Suzuki. J and Oba. K, "A Model-Driven Development Framework for Non-Functional Aspects in Service Oriented Grids" ICAS, IEEE Computer Society, pp 30-38, 2006

[10] S. Paunov, J. Hill, D. C. Schmidt, J. Slaby, and S. Baker, "Domain-Specific Modeling Languages for Configuring and Evaluating Enterprise DRE System Quality of Service". Proceedings of IEEE International symposium and Workshop on the Engineering of Computer Based Systems, pp 198-208, 2006

[11] D. C. Schmidt, "Model-Driven Engineering", IEEE Computer, 39(2), pp 25-31, 2006.

[12] Z. Gu, B. Xu, J. Li, "Inheritance-Aware Document- Driven Service Composition", Proceeding of IEEE International Conference on E-Commerce Technology and on Enterprise Computing, ECommerce, and E-Services, pp. 513-516, 2007.

[13] S.C. Oh, J.W. Yoo, H. Kil, D. Lee, and S. Kumara, "Semantic Web-Service Discovery and Composition Using Flexible Parameter Matching", Proceedings of IEEE International Conference on E-Commerce Technology and on Enterprise Computing, ECommerce, and E-Services, pp. 533-536, 2007.

[14] John Jung, Soundar Kumara, Dongwon Lee, and Seog, "A Web Service Composition Framework Using Integer Programming with Non-Functional Objectives and Constraints" IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, pp 347-350, 2008

[15] Asoke K Talukder and Manish Chaitanya, "Architecting Secure Software Systems", Auerbach Publications, 2008.

[16] Asoke K Talukder "Analyzing and Reducing the Attack Surface for a Cloud-ready Application" Indo-US Conference on Cyber Security, Cyber Crime, and Cyber Forensics, National Institute of Technology Karnataka, 2009

[17] G. Sindre and A.L. Opdahl, "Eliciting Security Requirements by Misuse Cases," in Proceedings of 37th Conference on Techniques of Object-Oriented Languages and Systems, TOOLS Pacific 2000, pp. 120–131, 2000

[18] A. Avizienis and J. C. Laprie. Dependable computing: from concepts to design diversity. In Proc. IEEE, 74(5):629–638, May 1986.

[19] www.issco.unige.ch

[20] J.C. C. Laprie, A. Avizienis, H. Kopetz. Dependability: Basic Concepts and Terminology. Springer-Verlag New York, 1992

[21] E. Thomas. Service-Oriented Architecture: Concepts, Technology, and Design. Prentice Hall, 2005.

[22] E. Newcomer. Understanding Web Services: XML, WSDL, SOAP, and UDDI. Addison-Wesley, 2002.

[23] G. Alonso, F. Casati, H. Kuno, and V. Machiraju. Web Services: Concepts, Architectures and Applications. Springer-Verlag, 2004.

[24] C. Peltz. Web services orchestration and choreography. Computer, 36(10):46–52, 2003.

[25] Heather, Hinton, Maryann Hondo, Beth Hurchison, "Security patterns within a Service Oriented Architecture", IBM, 2006.

[26] Paul Kearney, "Message Level Security for Web Service", Information Security Technical Report, Elsevier, Vol. 10, No. 1, 2005, pp 41-50

AUTHORS PROFILE

Saleem Basha is a Ph.D research scholar in the Department of Computer Science, Pondicherry University. He has obtained B.E in the field of Electrical and Electronics Engineering, Bangalore University, Bangalore, India and M.E in the field of Computer Science and Engineering, Anna University, Chennai, India. He is currently working in the area of web service modelling systems.

Dr. Dhavachelvan Ponnurangam is working as Associate Professor, Department of Computer Science, Pondicherry University, India. He has obtained his M.E. and Ph.D. in the field of Computer Science and Engineering in Anna University, Chennai, India. He is having more than a decade of experience as an academician and his research areas include Software Engineering and Standards, web service computing and technologies. He has published around 50 research papers in National and International Journals and Conferences. He is collaborating and coordinating with the research groups working towards to develop the standards for Attributes Specific SDLC Models & Web Services computing and technologies.

Data mining Aided Proficient approach for optimal inventory control in supply chain management

Chitriki Thotappa

Assistant Professor, Department of Mechanical Engineering,
Proudadevaraya Institute of Technology, Hospet.
Visvesvaraya Technological University, Karnataka, India
thotappa@gmail.com

Dr. Karnam Ravindranath

Principal
Annamacharya Institute of Technology, Tirupati
kravi1949@yahoo.com

Abstract— **Optimal inventory control is one of the significant tasks in supply chain management. The optimal inventory control methodologies intend to reduce the supply chain (SC) cost by controlling the inventory in an effective manner, such that, the SC members will not be affected by surplus as well as shortage of inventory. In this paper, we propose an efficient approach that effectively utilizes the data mining concepts as well as genetic algorithm for optimal inventory control. The proposed approach consists of two major functions, mining association rules for inventory and selecting SC cost-impact rules. Firstly, the association rules are mined from EMA-based inventory data, which is determined from the original historical data. Apriori, a classic data mining algorithm is utilized for mining association rules from EMA-based inventory data. Secondly, with the aid of genetic algorithm, SC cost-impact rules are selected for every SC member. The obtained SC cost-impact rules will possibly signify the future state of inventory in any SC member. Moreover, the level of holding or reducing the inventory can be determined from the SC cost-impact rules. Thus, the SC cost-impact rules that are derived using the proposed approach greatly facilitate optimal inventory control and hence make the supply chain management more effective.**

Keywords-SC cost; SC cost-impact rule; EMA-based inventory; Apriori; Genetic Algorithm (GA).

I. INTRODUCTION

Nowadays, supply chains are at the center stage of business performance of manufacturing and service enterprises [5]. A SC consists of all parties involved directly or indirectly and in satisfying a customer request. It includes suppliers, manufacturers, distributors, warehouses, retailers and even customers themselves [6]. Because of the intrinsic complexity of decision making in supply chains, there is a growing need for modeling methodologies, which help to identify and innovate strategies for designing high performance SC networks [5]. Research on supply chains makes an attractive field of study, offering several approach roads to organizational integration processes. Some of the problems are considered as most important, which canalize research project in the area of supply chains that are related to demand variability and demand distortion throughout the SC [7]. Modern supply chains are highly complex and dynamic; the number of facilities, the number of echelons, and the structure of material and information flow contribute to the complexity of the SC [9]. In addition, increases in the uncertainties in

supply and demand, globalization, reduction in product and technology life cycles, and the use of outsourcing in manufacturing, distribution and logistics resulting in more complex supply networks, can lead to higher exposure to risks in the SC [8].

The ultimate goal of every SC is to maximize the overall value generated by the chain, which depends on the ability of the organization to fulfill customer orders faster and more efficiently [9]. While the separation of SC activities among different companies enables specialization and economies of scale, many important issues and problems need to be resolved for a successful SC operation which is the main purpose of supply chain management (SCM) [14]. SCM is a traditional management tool [1] which has attracted increasing attention in the academic community and in companies looking for practical ways to improve their competitive position in the global market [4]. SCM is an integrated approach to plan and control materials and information flows [3]. Successful SCM incorporates extensive coordination among multiple functions and independent companies working together to deliver a product or service to end consumers [2]. Inventory control has been considered as a vital problem in the SCM for several decades [10].

Inventory is defined as the collection of items stored by an enterprise for future use and a set of procedures called inventory systems assist in examination and control of the inventory. The inventory system supports the estimation of amount of each item to be stored, when the low items should be restocked and the number of items that must be ordered or manufactured as soon as restocking becomes essential. The SC cost was hugely influenced by the overload or shortage of inventories [11]. Since inventory is one of the major factors that affect the performance of SC system, the effective reduction of inventory can substantially reduce the cost level of the total SC [13]. Thus, inventory optimization has emerged as one of the most recent topics as far as SCM is considered [11]. Inventory optimization application organizes the latest techniques and technologies, thereby assisting the enhancement of inventory control and its management across an extended supply network. Some of the design objectives of inventory optimization are to optimize inventory strategies, and thus used in enhancing customer service, reducing lead times and costs and meeting market demand [11].

Under the influence of the SCM, conventional inventory control theories and methods are no longer adapted to the new environment [12]. The optimal inventory control methodologies intended to reduce the SC cost in the SC network. They minimize the SC cost by controlling the inventory in an optimal manner and so that the SC members will not be affected by surplus as well as shortage of inventory. In order to control the inventory in an optimal manner, we propose an efficient approach with the effective utilization of data mining concepts as well as GA. The rest of the paper is organized as follows. Section II gives a brief introduction about the data mining and generating association rules using Apriori and Section III reviews some of the recent related works. Section IV details the proposed approach for optimal inventory control with required mathematical formulations. Section V discusses about the implementation results and Section VI concludes the paper.

II. DATA MINING

Data mining is one of the newly emerging fields, which is concerning the three worlds of Databases, Artificial Intelligence and Statistics. The information age has enabled several organizations in order to gather huge volumes of data. But, the utility of this data is negligible if "meaningful information" or "knowledge" cannot be extracted from it [15]. Data mining has been emerging as an effective solution to analyze and extract hidden potential information from huge volume of data. The term data mining is used for techniques and algorithms that allow analyzing data in order to determine rules and patterns describing the characteristic properties of that particular data. [21].

Usually, data mining tasks can be categorized into either prediction or description [18]. Clustering, Association Rule Mining (ARM) [19] and Sequential pattern mining are few descriptive mining techniques. The predictive mining techniques involve tasks like Classification [20], Regression and Deviation detection [34]. Data mining is utilized in both the private and public sectors. Data mining is usually used by business intelligence organizations, financial analysts and also used for healthcare management or medical diagnosis to extract information from the enormous data sets generated by modern experimental and observational methods [16] [17].

Generally, Data mining is used to extract interesting knowledge that can be represented in several various techniques such as clusters, decision trees, decision rules and much more. In these, association rules have been proved to be effective in identifying interesting relations in massive data quantities [25]. Association Rule Mining (ARM), initially introduced by Agrawal et al. [35], is a well-known data mining research field [24]. ARM correlates a set of items with other sets of items in a database [23]. It aspires to mine interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories [22]. ARM has a extensive range of applications in the fields of Market basket analysis, Medical diagnosis/ research, Website navigation analysis, Homeland security and so on [26]. ARM is to identify the association rule which satisfies the pre- defined minimum support and confidence from a given database. Association rule mining is a two step process [27]:

- Finding those itemsets whose occurrences exceed a predefined threshold in the database, these itemsets are called frequent or large itemsets.

- Generating association rules from those large itemsets with the constraints of minimal confidence.

The basic problem in mining association rules is mining frequent itemsets [30]. Frequent item set mining problem has received a great deal of attention [28] from its introduction in 1993 by Agarwal et al [35]. Frequent item sets play an significant role in several data mining tasks that tries to determine interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers, clusters and much more [29]. There have been several various algorithms developed for mining of frequent patterns, which can be classified into two categories. The first category, candidate-generation-and test approach, such as Apriori and second category of methods includes FP-growth and Tree Projection [30].

Apriori is one of the most popular data mining approaches for determining frequent itemsets from transactional datasets. The Apriori algorithm is the key basis of several other well-known algorithms and implementations [31]. The Apriori algorithm uses two values for rule construction: 1.) a support value and 2.) a confidence value. Depending on the setting of each index threshold, the search space can be reduced, or the candidate number of association rules can be increased. However, experience is necessary for setting an effective threshold [32]. The basic idea of Apriori algorithm is to generate a specific size of the candidate projects set, and then scan the database time's line counts, to determine whether the candidate frequent item sets [33].

III. RELATED WORKS

Some of the recent research works available in the literature are described in this section. A. L. Symeonidis et al., [36] have introduced a successful paradigm for coupling Intelligent Agent technology with Data Mining. Considering the state-of-the-art Multi-Agent Systems (MAS) development and SCM evaluation practices, they have proposed a methodology to identify the appropriate metrics for DM-enhanced MAS for SCM and used those metrics to evaluate its performance. They have also provided an extensive analysis of the methods in which DM could be employed to improve the intelligence of an agent, agent Mertacor. A number of metrics were applied to evaluate their results before incorporating the selected model with their agent. Their mechanism proved that their agent was capable of increasing its revenue by adjusting its bidding strategy.

Steven Prestwich et al. [37] have described a simple re-sampling technique called Greedy Average Sampling for steady-state GAs such as GENITOR. It requires an extra runtime parameter to be tuned, but does not need a large population or assumptions on noise distributions. While experimented on a well-known Inventory Control problem, it performed a large number of samples on the best

chromosomes yet only a small number on average, and was more effective than the other four tested techniques.

Mouhib Al-Noukari et al [38] have explained a data mining application in car manufacturing domain and experimented it. Their application results demonstrated the capability of data mining techniques in providing important analysis such as launch analysis and slow turning analysis. Such analysis helped in providing car market with base for more accurate prediction of future market demand.

Tao Ku et al. [39] have presented a complex event mining network (CEMN) and defined the fundamentals of radio-frequency identification (RFID)-enabled SC event management. Also, they have discussed how a complex event processing (CEP) could be used to resolve the underlying architecture challenges and complexities of integrating real-time decision support into the supply chain. Finally, a distributed complex event detection algorithm based on master-workers pattern was proposed to detect complex events and trigger correlation actions. Their results showed that their approach was more robust and scaleable in large-scale RFID application.

Se Hun Lim [40] has developed a control model of SCM sustainable collaboration using Decision Tree Algorithms (DTA). He has used logistic regression analysis (LRA) and multivariate determinate analysis (MDA) as a benchmark and compared the performance of forecasting SCM sustainable collaboration through three types of models LRA, MDA, DTA. Forecasting SCM sustainable collaboration using DTA was considered as the most outstanding feature. The obtained result has provided useful information of SCM sustainable collaboration determining factors in the manufacturing and distributing companies.

Shu-Hsien Liao et al. [41] have investigated functionalities that best fit the consumer's needs and wants for life insurance products by extracting specific knowledge patterns and rules from consumers and their demand chain. They have used the apriori algorithm and clustering analysis as methodologies for data mining. Knowledge extraction from data mining results was illustrated as market segments and demand chain analysis on life insurance market in Taiwan in order to propose suggestions and solutions to the insurance firms for new product development and marketing.

Xu Xu et al. [42] have proposed an approach that combines expert domain knowledge with Apriori algorithm to discover the pattern of supplier under the methodology of Domain-Driven Data Mining (D3M). Apriori algorithm of data mining with the help of Intuitionistic Fuzzy Set Theory (IFST) was employed during the process of mining. The obtained overall patterns help in deciding the final selection of suppliers. Finally, AHP was used to efficiently tackle both quantitative and qualitative decision factors involved in ranking of suppliers with the help of achieved pattern. An example searching for pattern of supplier was used to demonstrate the effective implementation procedure of their method. Their method could provide the guidelines for the decision makers to effectively select their suppliers in the current competitive business scenario.

IV. THE PROPOSED APPROACH FOR OPTIMAL INVENTORY CONTROL

In the proposed approach for optimal inventory control, two major functions are included, namely, association rules mining for inventory and recognizing optimal inventory rules to be maintained. Prior to perform the two aforesaid functions, a database of historical data has to be maintained. The database holds the historical record of inventory over N_p periods in N_S SC members, say, $[I_{ij}]_{N_P \times N_S}$; $0 \le i \le N_P - 1$ and $0 \le j \le N_S - 1$. Initially, the Exponential Moving Average (EMA) is determined for the historical data as follows

$$I_{ema_{lj}} = I_{prev_{lj}} + \alpha(I_{(l+n)j} - I_{prev_{lj}}) \; ; \; 0 \le l \le N_P - (n+1) \quad (1)$$

where,

$$I_{prev_{lj}} = \begin{cases} I_{ema_{(l-1)j}} & ; if \; l > 0 \\ \dfrac{1}{n}\sum_{i=0}^{n-1} I_{ij} & ; otherwise \end{cases} \quad (2)$$

The EMA values of the original historical data for $N_P - n$ periods, $[I_{ema_{lj}}]_{(N_P-n) \times N_S}$ from (1), where, $\alpha = 2/(n+1)$ (termed as constant smoothing factor), is subjected for a decision making process as follows

$$I'_{ema_{lj}} = \begin{cases} shortage & ; I_{ema_{lj}} < I_{th} \\ balance & ; I_{ema_{lj}} = I_{th} \\ excess & ; I_{ema_{lj}} > I_{th} \end{cases} \quad (3)$$

As given above, EMA-based inventory data $[I'_{ema_{lj}}]_{(N_P-n) \times N_S}$ is obtained in which the original historical data is converted into three different states of inventory which include shortage, balance and excess. Subsequently, the association rules for inventory are mined from the previously obtained EMA-based inventory data.

A. Mining Association rules for inventory using Apriori

One of the two major functions of the approach, mining association rules for inventory is described in the sub-section. Mining the association rules for inventory is to find the relationship between the inventories of the SC members. In the proposed approach, we utilize Apriori, a classic algorithm for learning the association rules. Let, $\left\{ I'_{ema_{l1}}, I'_{ema_{l2}}, I'_{ema_{l3}}, \cdots, I'_{ema_{lN_S}} \right\}$ be the itemset taken from the EMA-based inventory data $\left\{ I'_{ema_{lj}} \right\}_{(N_P-n) \times N_S}$. The itemset

and the dataset $\left\{I'_{ema_{lj}}\right\}_{(N_P-n)\times N_S}$ are subjected to Apriori for mining association rules. Initially, the Apriori finds the frequent itemsets with a minimum support threshold $s_{\min}$, and determines the rule which states the probabilistic relationship between the items in the frequent itemsets with a minimum confidence of $c_{\min}$.

The Apriori determines the association rules from the frequent itemset by calculating the possibility of an item to be present in the frequent itemset, given another item or items is present. For instance, considering a frequent itemset, $I'_{ema_{l1}}$, $I'_{ema_{l2}}$ and $I'_{ema_{l3}}$ in which a rule may be derived as when the inventory in $I'_{ema_{l1}}$ and $I'_{ema_{l2}}$ are excess in a period l; $l \in (0, N_P - (n+1))$, then the inventory in $I'_{ema_{l3}}$ is likely to be shortage. The general syntax of the rule for the aforesaid example is given as $(I'_{ema_{l1}} = \text{excess}, I'_{ema_{l2}} = \text{excess}) \rightarrow (I'_{ema_{l3}} = \text{Shortage})$; $c \geq c_{\min}$. Hence, by using the apriori, the association rules are mined with a minimal confidence $c_{\min}$ based on the frequent itemset with a minimal support $s_{\min}$.

The mined rules are given as $\{A\}_q \rightarrow \{B\}_q$; $0 \leq q \leq N_r - 1$, where, $\{A\}_q$ and $\{B\}_q$ are the antecedent and consequent of the qth rule respectively and N_r be the number of association rules generated. The antecedent and consequent consists of one or more items that belongs to the itemset $\left\{I'_{ema_{lj}}\right\}$ (i.e. $\{A\}_q \subseteq \left\{I'_{ema_{lj}}\right\}$, $\{B\}_q \subseteq \left\{I'_{ema_{lj}}\right\}$) and also it satisfies $\{A\}_q \cap \{B\}_q = \phi$. After obtaining the association rules, they are allocated for j^{th} SC member based on the consequent of the rules. The final rules after allocation are obtained as follows

$$R'_j = R_j - \phi \qquad (4)$$

where,

$$R_j = \begin{cases} \{A\}_q \rightarrow \{B\}_q; & if \ I'_{ema_{lj}} \in \{B\}_q \\ \phi & ; \ else \end{cases} \qquad (5)$$

Using (5), the rules R'_j which have the element $I'_{ema_{lj}}$ in the consequent are assigned to the jth SC member. Each SC member has its own rules that illustrate its inventory's state with respect to other SC member or members. So, N_S set of rules are obtained where each set has $\left|R'_j\right|$ number of rules and they need not to be in equal number. From the N_S set of rules, a rule per each SC member (i.e. a rule per set) is selected using GA. The rules are chosen in such a way that they have major impact over the SC cost.

B. Selecting SC cost-impact rules using GA

The obtained rules from apriori are the frequently occurred events in the past and so they illustrate that they have a good impact over the SC cost, but not strongly. To identify the rules that have strong impact over the SC cost (SC cost-impact rules), it is essential to consider the shortage cost and holding cost. It is already known that the SC cost increases, when either of the shortage and holding costs increases. Hence, by considering the shortage or holding cost in the GA, SC cost-impact rules can be obtained. The process of selecting SC cost-impact rules using GA is explained as follows

Step 1: Generate initial chromosomes, $X_a = [x_0^{(a)} \ x_1^{(a)} \ x_2^{(a)} \ \cdots \ x_{N_S-1}^{(a)}]$; $0 \leq a \leq N_{pop} - 1$, where N_{pop} is the population size. The jth gene of the chromosome $x_j^{(a)}$; $0 \leq j \leq N_S - 1$ is an arbitrary integer in the interval $(0, |R'_j| - 1)$, where, $\left|R'_j\right|$ is the cardinality of the rule set belongs to the jth SC member.

Step 2: Determine fitness of the chromosomes present in the population pool using the fitness function

$$f(a) = \cfrac{1}{\sum\limits_{j=0}^{N_S-1}\left(C_{I_j} \times \left|\mu_{ema_j}(R'_j(x_j^{(a)}))\right| \times c_{R'_j(x_j^{(a)})}\right)} \qquad (6)$$

where,

$$C_{I_j} = \begin{cases} S_{c_j}; & if \ \mu_{ema_j}(R'_j(x_j^{(a)})) < 0 \\ H_{c_j}; & if \ \mu_{ema_j}(R'_j(x_j^{(a)})) > 0 \\ 0 \ ; & if \ \mu_{ema_j}(R'_j(x_j^{(a)})) = 0 \end{cases} \qquad (7)$$

$$\mu_{ema_j}(R'_j(x_j^{(a)})) = \frac{1}{F_{R'_j(x_j^{(a)})}} \sum\limits_{k \in (0, N_P-(n+1))} I_{ema_{kj}} \qquad (8)$$

In (6), $f(a)$ is the fitness value of ath chromosome, C_{I_j} (determined using (7)) is the inventory cost incurred by the jth SC member, $\mu_{ema_j}(R'_j(x_j^{(a)}))$ (determined using (8)) is the mean EMA value of the $I_{ema_{lj}}$ that are taken only from

the pattern which satisfies the rule $R'_j(x_j^{(a)})$ and $c_{R'_j(x_j^{(a)})}$ is the confidence of the rule $R'_j(x_j^{(a)})$. In (7), S_{c_j} is the shortage cost incurred for a unit of shortage in j^{th} SC member, H_{c_j} is the holding cost incurred for a unit to hold in the jth SC member. In (8), $F_{R'_j(x_j^{(a)})}$ is the frequency of occurrence of data pattern that satisfies the rule $R'_j(x_j^{(a)})$ and $I_{ema_{k_j}}$ is the EMA value of inventory in j^{th} SC member that are available in the data pattern, where, $I_{ema_{k_j}} \in \left\{ I_{ema_{l_j}} \right\}$.

Step 3: Select the best $N_{pop}/2$ chromosomes, which have minimum fitness, from the population pool.

Step 4: Crossover the selected chromosomes with a crossover rate of CR so as to obtain $N_{pop}/2$ children chromosomes.

Step 5: Mutate the children with a mutation rate of MR which leads to $N_{pop}/2$ new chromosomes.

Step 6: Place the $N_{pop}/2$ new chromosomes and $N_{pop}/2$ parent chromosomes in the population pool.

Step 7: Go to step 2, until the process reaches a maximum number of iterations N_g. Once the process reaches N_g, terminate it and select the $N_{pop}/2$ best chromosomes, which have minimum fitness value.

The best chromosomes obtained from the GA indicate $N_{pop}/2$ set of rules in which each set has N_s rules (one rule per SC member). From the rule obtained for a particular SC member, it can be decided that

- The inventory will likely to be as in the rule given for the inventory of the associated SC members.

- Either by reducing or by increasing the holding level of inventory (can be decided from the rule) in the SC member, an optimal level of inventory can be maintained in the upcoming days.

Hence, by the optimal inventory control, the SC member will not be suffered either by increased shortage cost or by increased holding cost. This ultimately helps to keep the SC cost in a controlled manner.

C. Evaluation of Rules

The efficacy of the rules is demonstrated by comparing the obtained rules with all the remaining rules. To accomplish this, the SC cost and the confidence of the rule associated to the best chromosome are determined as

$$SC^{best} = \sum_{j=0}^{N_S - 1} \left(C_{I_j} \times \left| \mu_{ema}(R'_j(x_j^{best})) \right| \right) \qquad (9)$$

$$c^{best} = \frac{1}{N_S} \sum_{j=0}^{N_S - 1} c_{R_j(x_j^{best})} \qquad (10)$$

Similarly, the mean SC cost and the mean confidence are determined for all the remaining rules in the rule set $\left\{ R'_j \right\}$. Then, the efficacy is compared by determining the difference between the SC cost and confidence of the final SC cost-impact rule and the mean SC cost and the mean confidence of the remaining rules, respectively.

V. RESULTS AND DISCUSSION

The proposed approach for optimal inventory control has been implemented in the working platform of JAVA (version JDK 1.6) and the results are discussed in this section. The inventory data (weekly data) has been simulated for five years (i.e. $N_p = 260$) by considering five SC members (i.e. $N_s = 5$), an agent A_1 and four retailers, R_1, R_2, R_3 and R_4. In the simulated inventory data, the negative and positive values represent the shortage amount of inventory and excess amount of inventory respectively. All the SC members have been considered to have the shortage cost and holding cost as $S_c = Rs.2.50$ and $H_c = Rs.1.00$ respectively. The I'_{ema} determined from the simulated data with $n = 7$ is given in the Table I.

TABLE I. A SAMPLE OF EMA-BASED INVENTORY DETERMINED FROM THE SIMULATED DATA

Sl. No	A₁	R₁	R₂	R₃	R₄
1	Excess	Shortage	Excess	Excess	Excess
2	Excess	Shortage	Excess	Shortage	Shortage
3	Excess	Shortage	Excess	Shortage	Shortage
4	Excess	Shortage	Excess	Shortage	Excess

The first major function of the proposed approach, mining association rules for inventory using Apriori has been implemented with the aid of data mining software WEKA (version 3.7). The Table II and the Table III consist of some frequent itemsets with $s_{min} = 10\%$ that are discovered from the I'_{ema} and some of the association rules generated from the discovered frequent itemset respectively. The rules that are categorized based on the consequent are shown in the Table IV.

TABLE II. SOME FREQUENT ITEMSETS DISCOVERED FROM I'_{ema} AT LENGTHS L_1, L_2, L_3 AND L_4, AND THEIR SUPPORT.

Length of the itemset	Frequent itemset	Support %
L_1	R_1=Shortage 0.536	53.6
	R_1=Excess 0.476	47.6
	R_2=Excess 0.6	60
L_2	R_1=Shortage, R_2=Excess 0.316	31.6
	R_1=Shortage, R_2=Shortage 0.22	22
	R_1=Shortage, R_3=Excess 0.224	22.4
L_3	R_1=Shortage, R_2=Excess, R_3=Excess 0.128	12.8
	R_1=Shortage, R_2=Excess, R_3=Shortage 0.188	18.8
	R_1=Shortage, R_2=Excess, R_4=Excess 0.132	13.2
L_4	R_1=Shortage, R_2=Excess, R_3=Shortage, R_4=Shortage 0.1	10
	R_1=Shortage, R_2=Excess, R_3=Shortage, A_1=Shortage 0.104	10.4

TABLE III. SOME GENERATED ASSOCIATION RULES WITH $c_{min} = 30\%$ AND THEIR CONFIDENCE

Sl. No	Association Rules	Confidence %
1	R_2=Excess, R_4=Shortage, A=Excess ==> R_1=Shortage	79
2	R_1=Excess, A=Excess ==> R_3=Shortage	76
3	R_1=Excess, R_4=Excess, A=Shortage ==> R_2=Excess	75
4	R_1=Shortage, R_4=Shortage, A=Excess ==> R_2=Excess	72

TABLE IV. SOME OF THE RULES THAT ARE CATEGORIZED BASED ON THE CONSEQUENT OF THE RULES

Sl. No	Rule for A_1	Rule for R_1	Rule for R_2	Rule for R_3	Rule for R_4
1	(R_1=Excess, R_3=Excess) → A1=Shortage	(R_2=Excess, R_4=Shortage, A1=Excess) → R_1=Shortage	(R_1=Excess, R_4=Excess, A1=Shortage) → R_2=Excess	(R_1=Excess, A1=Excess) → R_3=Shortage	(R_1=Excess ,R_2=Excess, R_3=Shortage) → R_4=Excess
2	(R_2=Shortage, R_3=Excess) → A1=Shortage	(R_3=Excess, A1=Excess) → R_1=Shortage	(R_1=Shortage, R_4=Shortage, A1=Excess) → R_2=Excess	(R_4=Excess, A1=Excess) → R_3=Shortage	(R_1=Excess, R_2=Excess) → R_4=Excess
3	(R_1=Shortage, R_2=Shortage) → A1=Shortage	(R_3=Excess, R_4=Shortage) → R_1=Shortage	(R_1=Excess, R_4=Excess) → R_2=Excess	(R_2=Shortage, A1=Excess) → R_3=Shortage	(R_1=Shortage ,R_3=Excess) → R_4=Shortage
4	(R_2=Shortage ,R_4=Shortage)→ A1=Shortage	(R_2=Excess, R_4=Shortage) → R_1=Shortage	(R_1=Excess, R_3=Shortage, R_4=Excess) → R_2=Excess	(R_1=Shortage, R_2=Excess, A1=Shortage) → R_3=Shortage	(R_2=Shortage, R_3=Excess) → R_4=Shortage

In selecting the SC cost-impact rules, the GA has been initialized with a chromosome length = 5 (i.e. *number of genes* = 5), $N_{pop} = 10$ and $N_g = 50$. The generated initial chromosome and the rules that are associated to the chromosome are given in the Fig. 1 and the Table V, respectively.

0	1	2	3	4
78	58	86	59	17

Figure 1. An initial chromosome of length '5' with random values in their genes

TABLE V. THE RULES ASSOCIATED TO THE CHROMOSOME, WHICH IS GIVEN IN THE FIG. 1.

Gene. no	Associated rules
0	R_4 = -12.46 --> R_1 = -11.7, R_3 = -8.64
1	R_3 = -11.32 --> R_2 = -9.09
2	R_4 = -12.46 --> R_1 = -11.7, R_3 = -8.64
3	R_2 = -8.84 --> R_4= 6.26
4	R_1 = -11.91 --> A_1 = -45.6

The generated chromosomes have been subjected to crossover with $CR = 0.6$ and the obtained children have been

subjected to mutation with $MR = 0.4$. In the mutation, the gene values in the mutation point are changed arbitrarily so that new chromosome is obtained from the child chromosome.

The final SC cost-impact rules that are associated to the obtained best chromosomes are given in the Table VI.

TABLE VI. SOME OF THE FINAL SC COST-IMPACT RULES ASSOCIATED TO THE BEST CHROMOSOMES OBTAINED FROM GA.

Solution no.	Best SC cost-impact Rules				
	A_1	R_1	R_2	R_3	R_4
1	$R_4 = -11.84 \rightarrow R_2 = -10.89$, $A_1 = -45.89$	$R_4 = -10.33 \rightarrow R_1 = -12.02$, $A_1 = -46.68$	$R_4 = -11.62 \rightarrow R_2 = -10.02$	$R_4 = 9.67 \rightarrow R_1 = 14.12$, $R_3 = -12.52$	$R_3 = -8.55 \rightarrow R_4 = -12.11$
2	$R_4 = -11.84 \rightarrow R_2 = -10.89$, $A_1 = -45.89$	$(R_2 = 10.15, A_1 = 25.94) \rightarrow R_1 = -11.56$	$R_4 = -11.62 \rightarrow R_2 = -10.02$	$R_4 = 9.67 \rightarrow R_1 = 14.12$, $R_3 = -12.52$	$R_3 = -8.55 \rightarrow R_4 = -12.11$

All the obtained rules in a solution provide their combined contribution in the SC cost. The SC cost given by the solution was very high in the past records and so, by considering those rules in the solution, the SC cost can be reduced in the future. The cost reduction can be accomplished by inverse holding of inventory that has been obtained as a rule for a particular SC member. From Table VI, by keeping 46, 12, 10, 13 and 12 (approximately) units of products additionally in the SC member A_1, R_1, R_2, R_3 and R_4, respectively, the SC cost will be reduced in the future. For evaluation, the SC^{best}, c^{best} (from solution I), SC^{mean} SC mean, c^{mean} have been determined and tabulated in the Table VII.

TABLE VII. COMPARISON OF THE OBTAINED SC-COST IMPACT RULE AND THE REST OF THE RULES IN THE RULE SET $\left\{R'_j\right\}$ BASED ON THE SC COST AND THE CONFIDENCE OF THE RULE.

Sl. no.	Efficacy Factor	SC Cost-Impact rule	Rest of the Rule set $\left\{R'_j\right\}$
1	Total SC Cost (in Rs.)	222.50	145.40
2	Mean Confidence (in %)	51.93	38.60

From Table VII, it can be demonstrated that the SC cost-impact rule which is obtained from best chromosome claims more SC cost as well as more frequency of occurrence rather than the all the other rules. Hence, by considering the rule, the optimal inventory can be maintained in all the SC members and so SC can be reduced effectively.

VI. CONCLUSION

In the paper, an efficient approach for optimal inventory control using Apriori and GA has been proposed and implemented as well. For experimentation, we have utilized the EMA-based inventory data determined from the simulated data. The results have shown that the effectual association rules are mined from the EMA-based inventory data using Apriori. Then, the rules have been categorized based on their consequent, followed by the selection of SC cost-impact rules using GA. The fitness function devised for the GA has performed well in selecting the rules that have high impact on the SC cost. It could be decided that, the upcoming inventory in any SC member will likely be as in the obtained SC cost-impact rules. It could also be decided, whether the inventory has to be reduced or increased in the particular SC member. Also, an EMA level of inventory to be reduced or increased can also been determined from the obtained SC cost-impact rules. Thus, the SC cost will be reduced proficiently by the proposed optimal inventory control approach that paves the way for effective SCM.

REFERENCES

[1] Duangpun Kritchanchai and Thananya Wasusri, "Implementing Supply Chain Management in Thailand Textile Industry", International Journal of Information Systems for Logistics and Management, Vol.2, No.2, pp.107-116, 2007.

[2] Jennifer Blackhurst, Christopher W. Craighead and Robert B, "Towards supply chain collaboration: an operations audit of VMI initiatives in the electronics industry", Int. J. Integrated Supply Management, Vol. 2, No. 1/2, pp. 91-105, 2006.

[3] Shen-Lian Chung and Hui-Ming Wee, "Pricing Discount For A Supply Chain Coordination Policy With Price Dependent Demand", Journal of the Chinese Institute of Industrial Engineers, Vol. 23, No. 3, pp. 222-232, 2006.

[4] Xiande Zhao, Jinxing Xie and Janny Leung, "The impact of forecasting model selection on the value of information sharing in a supply chain", European Journal of Operational Research, Vol.142, pp.321–344, 2002.

[5] Shantanu Biswas And Y. Narahari, "Object oriented modeling and decision support for supply chains", European Journal of Operational Research, vol. 153, No. 3, pp. 704-726,2004.

[6] M. Zandieh and S. Molla- Alizadeh- Zavardehi, "Synchronized Production and Distribution Scheduling with Due Window", in proceedings of Journal on Applied Sciences, vol. 8, no. 15, pp: 2752- 2757, 2008.

[7] Francisco Campuzano Bolarín, Andrej Lisec and Francisco Cruz Lario Esteban, "Inventory Cost Consequences of Variability Demand Process within A Multi-Echelon Supply Chain", Journal of Logistics and Sustainable Transport, vol. 1, No.3, 2008.

[8] Vasco Sanchez Rodrigues, Damian Stantchev, Andrew Potter and Mohamed Naim and Anthony Whiteing, "Establishing a transport operation focused uncertainty model for the supply chain", International Journal of Physical Distribution & Logistics Management, Vol. 38 No. 5, pp. 388-411, April 2008.

[9] Mustafa Rawata and Tayfur Altiokb, "Analysis of Safety Stock Policies in De-centralized Supply Chains", International Journal of Production Research, Vol. 00, No. 00, pp. 1-22, March 2008.

[10] Mileff, Péter, Nehéz, Károly, "A new inventory control method for supply chain management", 12th International Conference on Machine Design and Production, 2006.

[11] P. Radhakrishnan, V.M. Prasad and M.R. Gopalan, "Optimizing Inventory Using Genetic Algorithm for Efficient Supply Chain Management," Journal of Computer Science, Vol. 5, No. 3, pp. 233-241, 2009.

[12] Guangyu Xiong and Hannu Koivisto, "Research on Fuzzy Inventory Control under Supply Chain Management Environment," in proceedings

of Applied Simulation and Modelling, pp. 907–916, September 3 – 5, Marbella, Spain, 2003.

[13] Guangshu Chang, "Supply Chain Inventory Level with Procurement Constraints", International Conference on Wireless Communications, Networking and Mobile Computing, 2007, WiCom 2007, p.p. 4931-4933, DOI. 10.1109/WICOM.2007.1208.

[14] Peter Trkman and Ales Groznik, "Measurement of Supply Chain Integration Benefits", Interdisciplinary Journal of Information, Knowledge, and Management Volume 1, p.p. 37-45, 2006.

[15] Yehuda Lindell and Benny Pinkas, "Privacy Preserving Data Mining", journal of Cryptography, vol. 15, no. 3,2002.

[16] David L. Iverson, "Data Mining Applications for Space Mission Operations System Health Monitoring", NASA Ames Research Center, Moffett Field, California, 94035, 2008.

[17] Ping Lu, Brent M. Phares, Terry J. Wipf and Justin D. Doornink, "A Bridge Structural Health Monitoring and Data Mining System", in Proceedings of the 2007 Mid-Continent Transportation Research Symposium, Ames, Iowa, August 2007.

[18] Tibebe Beshah Tesema, Ajith Abraham And Crina Grosan, "Rule Mining And Classification of Road Traffic Accidents Using Adaptive Regression Trees", In Proc. Of I. J. On Simulation, Vol. 6, No. 10 and 11, 2008.

[19] F. Coenen, Leng, P., Goulbourne, G., "Tree Structures for Mining Association Rules", Journal of Data Mining and Knowledge Discovery, Vol 15, pp: 391-398, 2004.

[20] Hewen Tang, Wei Fang and Yongsheng Cao, "A simple method of classification with VCL components", proceedings of the 21st international CODATA Conference, 2008.

[21] Gerhard Münz, Sa Li, and Georg Carle, "Traffic anomaly detection using k-means clustering", in proceedings of GI/ITG-Workshop, Hamburg, Germany, September 2007.

[22] Sotiris Kotsiantis and Dimitris Kanellopoulos, "Association Rules Mining: A Recent Overview", GESTS International Transactions on Computer Science and Engineering, vol. 32, no. 1, pp: 71- 82, 2006.

[23] Huebner, Richard A., "Diversity- Based Interestingness Measures for Association Rule Mining", in proc. of ASBBS Annual Conference, vol. 16, no. 1, Feb. 2009.

[24] Yanbo J. Wang, Qin Xin and Frans Coenen, "Hybrid Rule Ordering in Classification Association Rule Mining", Transactions on Machine Leaning and Data Mining, vol. 1, no. 1, pp: 1- 15, 2008.

[25] Rahman AliMohammadzadeh, Sadegh Soltan and Masoud Rahgozar, "Template Guided Association Rule Mining from XML Documents", in proceedings of the 15th international conference on World Wide Web, pp: 963- 964, 2006.

[26] M.H.Margahny and A.A.Mitwaly, "Fast Algorithm for Mining Association Rules", in proc. of AIML Conference, 19- 21 December 2005.

[27] Kamrul Abedin Tarafder , Shah Mostafa Khaled , Mohammad Ashraful Islam , Khandakar Rafiqual Islam, Hasnain Feroze, Mohammed Khalaquzzaman and Abu Ahmed Ferdaus, "Reverse Apriori Algorithm for Frequent Pattern Mining", in proc. of Asian Journal on Information Technology, vol. 7, no. 12, pp: 524- 530, 2008.

[28] Bart Goethals, "Memory issues in frequent itemset mining", Proceedings of the 2004 ACM symposium on Applied computing, Nicosia, Cyprus, pp: 530-534, 2004.

[29] Bart Goethals, "Survey on Frequent Pattern Mining", Technical report, Helsinki Institute for Information Technology, 2003.

[30] M.H.Margahny and A.A.Shakour, "Scalable Algorithm for Mining Association Rules", in proc. of AIML Journal, vol. 6, no. 3, Sept. 2006.

[31] E. Ansari, G.H. Dastghaibifard, M. Keshtkaran and H.Kaabi, "Distributed Frequent Itemset Mining using Trie Data Structure", in proc. of Intl. Journal on Computer Science, vol. 35, no. 3, 21 August 2008.

[32] Ayahiko Niimi and Eiichiro Tazak, "Rule Discovery Technique Using Genetic Programming Combined with Apriori Algorithm", Lecture Notes In Computer Science, Springer-Verlag, vol. 273- 277, London, UK, 2000.

[33] Han Feng, Zhang Shu- Mao and Du Ying- Shuang, "The analysis and improvement of Apriori algorithm", in proc. of Journal on Communication and Computer, vol. 5, no. 9, Sept. 2008.

[34] S.Shankar, T.Purusothaman, "Utility Sentient Frequent Itemset Mining and Association Rule Mining: A Literature Survey and Comparative Study", International Journal of Soft Computing Applications, ISSN: 1453-2277 Issue 4 (2009), pp.81-95

[35] R. Agrawal, T. Imielinski, and A.N. Swami, "Mining association rules between sets of items in large databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207{216, ACM Press, 1993.

[36] L. Symeonidis, V. Nikolaidou and P. A. Mitkashave, "Sketching a methodology for efficient Supply chain management agents enhanced through Data mining", International Journal of Intelligent Information and Database Systems Vol. 2, Issue. 1, pp: 49-68, 2008.

[37] Steven Prestwich, S. Armagan Tarim, Roberto Rossi and Brahim Hnich, "A Steady-State Genetic Algorithm With Resampling for Noisy Inventory Control", Lecture Notes in Computer Science, Parallel Problem Solving from Nature – PPSN X, 2008.

[38] Mouhib Al-Noukari and Wael Al-Hussan, "Using Data Mining Techniques for Predicting Future Car market Demand," in proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications, pp. 1 - 5, 7-11 April, 2008.

[39] Tao Ku, YunLong Zhu and KunYuan Hu, "A Novel Complex Event Mining Network for Monitoring RFID-Enable Application," Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Vol. 2, pp. 925-929, 19-20 December, 2008

[40] Se Hun Lim, "The Design of Controls in Supply Chain Management Sustainable Collaboration Using Decision Tree Algorithm", in proc. of Intl. Journal on Computer Science and Network Security, vol. 6, no. 5A, May 2006.

[41] Shu-Hsien Liao, Ya- Ning Chen and Yu- Tia Tseng, "Mining demand chain knowledge of life insurance market for new product development", in proc. of Intl. Journal on Expert Systems with Applications, vol. 36, no. 5, pp: 9422- 9437, July 2009.

[42] Xu Xu, Jie Lin and Dongming Xu, "Mining pattern of supplier with the methodology of domain-driven data mining", in proc. of IEEE International Conference on Fuzzy System, pp: 1925- 1930, 20- 24 Aug., 2009.

Chitriki Thotappa received the B.E (Mech) and M.E (Production Management) degrees from the Department of Mechanical Engineering from Gulbarga and Karnataka Universities, Karnataka, INDIA in 1991 and 1994 respectively, he is currently pursuing the Ph.D. degree in the field of Supply Chain Management and closely working with his research supervisor **Dr. Karnam Ravindranath.** He is presently working as a Assistant Professor in the Department of Mechanical Engineering, Proudadevaraya Institute of Technology, Hospet. Visvesvaraya Technological University, Karnataka India and also visiting faculty for Diploma, and P.G Courses. And is a member for Professional bodies like MISTE and MIE.

Dr. Karnam Ravindranath received the B.E (Mech), M.E (Industrial Engg.) Degrees from Sri Vekateshwara University, Tirupati Andra Pradesh India in 1971 and 1976 respectively and later completed his Ph.D from Institute of Technology, Delhi in 1985. He worked as a Professor and Head, Department of Mechanical Engineering and also Principal of Sri Venkateshwara College of Engineering, Sri Venkateshwara University, Tirupati Andra Pradesh INDIA. During this period he has visited Pennsylvania University, **USA**, and Hamburg University, **Germany** and presented papers in International conference. He has awarded best teacher by the Govt. of Andhra Pradesh in 2007, and having a teaching experience of 32 years, he worked as a Dean faculty of Engineering, Chairman Board of Studies (UG and PG) and also Dean of Examinations. He has more than 70 research paper publications in International and National journals in his credit. He has produced 7 Ph.D scholars and another 8 are in pipeline. Presently working as a Principal of Annamacharya Institute of Technology, Tirupati. And is a member for Professional bodies like MISTE and MISME.

Robust Video Watermarking Algorithm Using Spatial Domain Against Geometric Attacks

Sadik Ali. M. Al-Taweel[1], Putra. Sumari[2], Saleh Ali K. Alomari[1,2]

[1, 2]School of Computer Science, Universiti Sains Malaysia

11800 Penang, Malaysia

sadiq_altaweel@yahoo.com, putras@cs.usm.my, salehalomari2005@yahoo.com

Abstract— **it is important for Digital watermarking to have digital data and multimedia, such as video, music, text, and image copyright protection because of network and multimedia techniques that easily copy. One of the significant problems in video watermarking is the Geometric attacks. In this paper new robust watermarking algorithm has been proposed, based on spatial domain which is robust against geometric attacks such as downscaling, cropping, rotation, and frame dropping. Besides, the embedded data rate is high and robust. The experimental results show that the embedded watermark is robust and invisible. The watermark was successfully extracted from the video after various attacks.**

Keywords-Video watermarking, geometric attacks, copyright protection.

I. INTRODUCTION

Digital watermarking has recently become a popular area of research due to the proliferation of digital data (image, audio, or video) in the Internet and the necessity to find a way to protect the copyright of these materials. Visible watermarks are visual patterns like logos, which are inserted into the digital data. Most watermarking systems involve marking imperceptible alteration on the cover data to convey the hidden information. This is called the invisible watermarks. Digital watermarks, on the other hand, are found with the advancement of the Internet and the ambiguity of digital data. Thus, it is natural to extend the idea of watermarking into the digital data. Recently, numerous digital watermarking algorithms have been developed to help protect the copyright of digital images and to verify the multimedia data integrity [1]. In spite of the existence of watermarking technique(s) for all kinds of digital data, most of the literatures address the watermarking of the still images for copyright protection and only some are extended to the temporal domain for the video watermarking [2],[3].

In this paper, we propose an oblivious video watermarking technique based on the spatial domain which is robust against geometric attacks, Besides, the embedded data rate is high and robust. This paper is organized as follows: Section 2 describes the related work; section 3 describes the proposed algorithm. Section 4 describes the performance evaluation.

II. RELATED WORK

Some of the video watermarking techniques targeting geometric attacks are on raw videos [4], [5]. Hartung and Girod proposed algorithm for uncompressed and compressed video watermarking, based on the idea of spreading the watermark energy over all of the pixels in each of the frames. The bit rate of the watermark is low, and it is not robust to frame loss [6].

Numerous video watermarking approaches suggested various ways of handling geometric attacks and they can be classified into several categories: invariant watermark [7], [8] synchronization [9], and autocorrelation [10].

Invariant watermarking embeds the watermark in a geometric-invariant transform, such as a log-polar wavelet transform, eliminating the need to identify and reverse the specific geometric attacks, such as rotation, and scaling. These kinds of techniques are very weak against a slight geometric distortion, such as small-angle rotation and near-one scaling. Moreover, the computational cost is too high to obtain the invariant domain from the varied transform.

The synchronization is the exhaustive search which entails inversion of a large number of possible attacks and testing for a watermark after each one. Since the number of possible attacks increases, the positive probability and computational cost become unacceptable.

The autocorrelation technique is similar to the synchronization approach. It spreads lots of extra data, in addition to the real watermark information to obtain synchronization for the watermark detection by autocorrelation, which either further distorts the host media or sacrifices the watermark payload.

Chan et al, presented a novel DWT-based video watermarking scheme with scrambled watermark and error correcting code [11]. The scheme is robust against attacks such as frame dropping, frame averaging, and statistical analysis. Campisi et al proposed the perceptual mask, applied in the 3D DWT domain and robust against MPEG2 and MPEG-4 compression, collusion and transcoding attacks [12].

Elbasi proposed a robust mpeg video watermarking in wavelet domain which is embedded in two bands (LL and

HH) and chosen attacks, JPEG compression, resizing, adding Gaussian noise and low pass filtering [13].

Anqiang presented adaptive watermarking scheme based on the error correction code and Human Visual System (HVS) in 3D-DWT domain. The proposed method is to resist the signal processing attacks, Gaussian noise, and frame dropping [14].

Xu Da-Wen proposed a method based on the 3D wavelet transforms. In this method, the original video frames are divided into 3D-blocks according to the HVS properties. The proposed method is robust against lossy compression; frame swapping, frame dropping and median filtering [15].

Al-Taweel and Sumari proposed video watermarking technique based on the DWT based on the spread spectrum communication. The proposed method is robust against JPEG compression, geometric attacks such as Downscaling, Cropping, and Rotation, as well as noising [16].

Al-Taweel and Sumari proposed video watermarking technique based on the discrete cosine transform domain based on the spread spectrum communication. The proposed method is robust against JPEG compression, geometric attacks such as downscaling, cropping, and Rotation, as well as noising such as guaussian noise and salt & pepper noise [17].

Al-Taweel and Sumari proposed a novel DWT-based video watermarking algorithm is proposed based on a three-level DWT using Haar filter which is robust against geometric distortions such as Downscaling, Cropping, and Rotation. It is also robust against Image processing attacks such as low pass filtering (LPF), Median filtering, and Weiner filtering. Furthermore, the algorithm is robust against Noise attacks such as Gaussian noise, Salt and Pepper attacks [18].

Essaouabi and Ibnelhaj presented video watermarking algorithm in the three-dimensional wavelet transform. The proposed algorithm is robust against the attacks of frame dropping, averaging and swapping [19].

The main significance of our technique is that it attempts to realize a good compromise between robustness performance, quality of the embedding and computational cost.

III. THE PROPOSED ALGORITHM

In order to meet the requirements of invisibility and robustness, an algorithm has been proposed that adaptively modifies the intensities of the host frames pixels, in such a manner that it is unnoticeable to human eyes. The proposed algorithm divides the host frame into a predefined number of blocks; it also modifies the intensities of the pixels depending on the contents of the blocks. For security requirements, private keys have also been used in this algorithm.

In this section, the overview of proposed watermarking scheme is shown in figure (1). The scheme is composed of four main components: watermark modulation, watermark embedding, frame dropping, and finally watermark extraction.
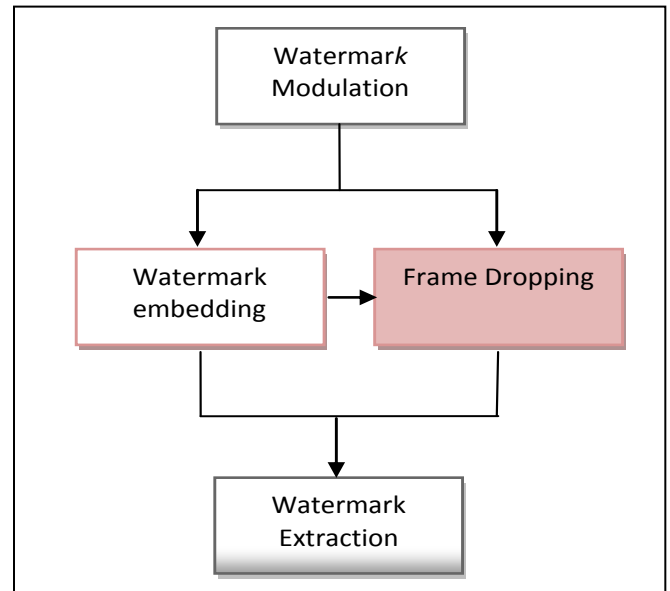
Figure 1. Model of watermarking algorithm

More details about these four main steps can be found in the next sections.

A. Watermark Modulation

The watermark L= [11,12,....,1N] with li $\in$ {0,1}, is a bit sequence of length N, which may be a meaningful image, like a logo of images of an owner.

The watermark is modulated by a bit-wise logical XOR operation, that contains a pseudo-random bit sequence s = [s1, s2, ..., sN] with si $\in$ {0,1} which is than multiplied by another pseudo-number sequence (0,1) to provide the modulated watermark sequence W = [w1, w2, …, wN], as shown in Figure (2).

The seed values of the two pseudo-random number generators are regarded as the two private keys for the proposed algorithm.
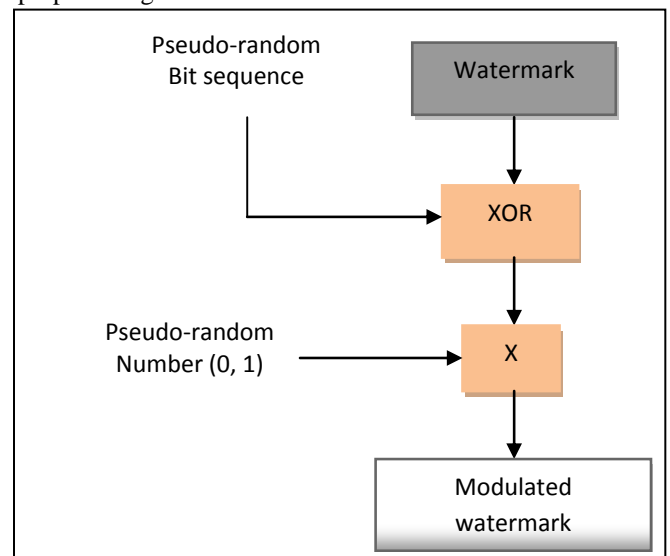
Figure 2. Modulation of the watermark

B.   *Watermark Embedding Process*

The modulated watermark bits are inserted into the host frames blocks, depending on the contrast of the block. Before the embedding process the host frame is decomposed into n×n blocks and the value of n is found as follows:

$$n = \sqrt{\frac{(M \times N)}{(X \times Y)}} \qquad (1)$$

Where M, N and X, Y represent the dimensions of the host image and the watermark respectively. The process of embedding in each block is carried out according to the following procedure.

1- Splitting the video into frames I, B, P
2- Calculate the mean, maximum and minimum values of the block.
3- Find the values in the block that are above and below the mean value.
4- Calculate the mean values of those below the blocks mean value and the mean values of those above it.
5- Calculate the new pixels values V$^{`}$ according to the following:

➢ Inserted bit 0
    If V< m_{low} then
        V'=V_{min}
    Else
        If V_{mean}<V<m_{high} then
            V$^{'}$$_{=}$$V_{mean}$
        Else
            V$^{'}$$_{=}$V-*a*

➢ Inserted bit >0
    If V<m_{high} then
        V'=V_{max}
    Else
        If m_{low}< V < V_{mean} then
            V$^{'}$$_{=}$$V_{mean}$
        Else
            V$^{'}$$_{=}$V+*a*

Where V is the original intensity, V_{mean}, V_{max}, V_{min} represent mean, maximum and minimum values of the blocks respectively. Whereas m_{low} and m_{high} represent the mean values of the pixels above and below the mean value of the block respectively.

6-   Finally the original frame is replaced with the resulting watermarked frame.

C.   *Watermark Extraction*

According to the embedding procedure, the sum of pixels in the watermarked block is larger than that of the original frame if the embedded bit is 1.

On the other hand, if the embedded bit is 0, than the sum of pixels in the watermarked block is smaller than that of the original frame. Hence, the original and the watermarked frames are used in the extraction process. Both of the frames are divided into the same blocks, which are used for the embedding process.

The sum of pixels for each corresponding block is computed, and if the sum of the original frame block pixels is greater than that of the watermarked frame, the extracted bit is considered to be 0, otherwise it is considered to be 1.

The extracted bits are then processed by XOR, with the same pseudo-random sequence used for embedding to produce the extracted watermark.

D.   *Watermarking Robust Against Frame Dropping*

The effect of cropping and downscaling is similar for each frame, whereas the frame dropping is unequal on less significant frames from the scenes of the video. For the embedded watermark to be robust against frame dropping, a proposed method has been illustrated in Figure. (3), where the original video is segmented into scenes, then the digital watermark is divided into a number of blocks according to the number of scenes. The goal of dividing the watermark is for embedding each block of watermark into its local scene (for more details Figure (4) and Figure (5) illustrate the embedding and extracting operations). Combining the technique mentioned in Section 3 will make the watermark robust against cropping, downscaling, rotation and frame dropping.
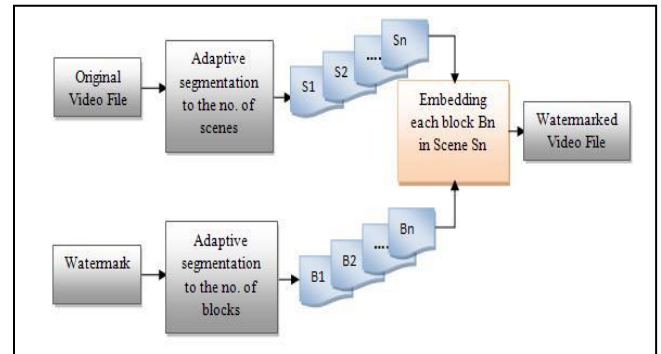
Figure 3. block diagram of proposed method for frame dropping

1)   *Embedding Watermark*

As shown in figure 5.7 the steps of embedding watermark against frame dropping as follow:

a)   Read watermark logo (modulated watermark).
b)   Segment watermark data into no of blocks according to the number of scenes.
c)   Embedded block no 1in the frames of the scene no 1.
d)   Embedded block no 2 in the frames of scene no 2
e)   Still embedded each block of watermark into its local scene.
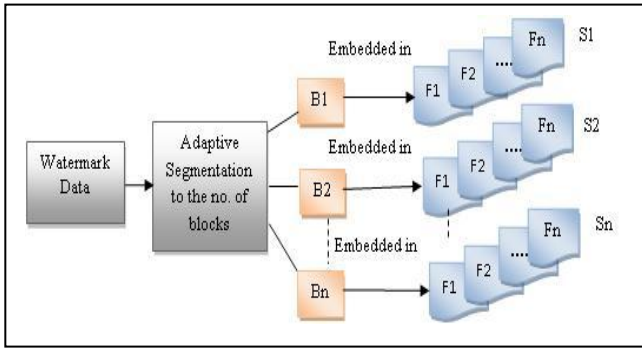
Figure 4. essential operation of embedding each block of watermark in scenes of video.

2) Extracting Watermark

As shown in figure 5.8 the steps of extracting watermark against frame dropping as follow:

a) Read watermarked video file.
b) Segmented the video into to the no of scenes.
c) Extracting each block from any frame of local scene.
d) Collect all extracted blocks.
e) Reconstruct watermark data.
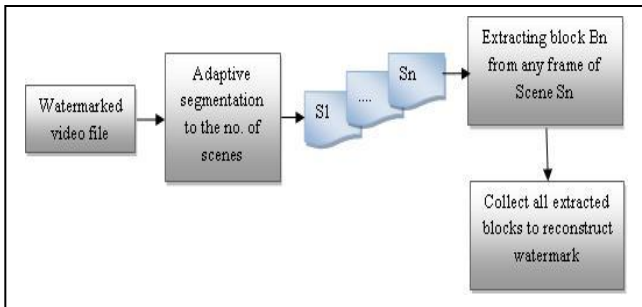
Figure 5. extracting algorithms

IV. PERFORMANCE EVALUATION

The proposed algorithm has been implemented in MATLAB version 7.5 and the experiments have been performed on a Pentium 4 PC running Windows XP. The performance of the proposed video watermarking algorithm has been evaluated on the basis of, imperceptibility and robustness. The metrics were evaluated using the standard video clips of 704×480 and 352x240 with size CIF and format 4.2.0 as shown in the table (6.1). A 64×64 binary logo (USM) shown in Figure (6), will also be embedded into this. In fact, experimental results indicate that the algorithm is very robust to geometric attacks. Figure (7) shows the original I-frame for test clips, watermarked frame for test clips and extracted watermark.

TABLE I.   VIDEO CLIPS USED IN TESTING

Video test sequence	Size	Format	Frames	Resolution
Susie on the phone	CIF	4.2.0	450	352×240
Flower garden	CIF	4.2.0	150	352×240
Football	CIF	4.2.0	150	704×480
Mobile and calendar	CIF	4.2.0	450	704×480
Tempete	CIF	4.2.0	149	352×288
Table Tennis	CIF	4.2.0	150	352x240

Figure.6. original watermark logo
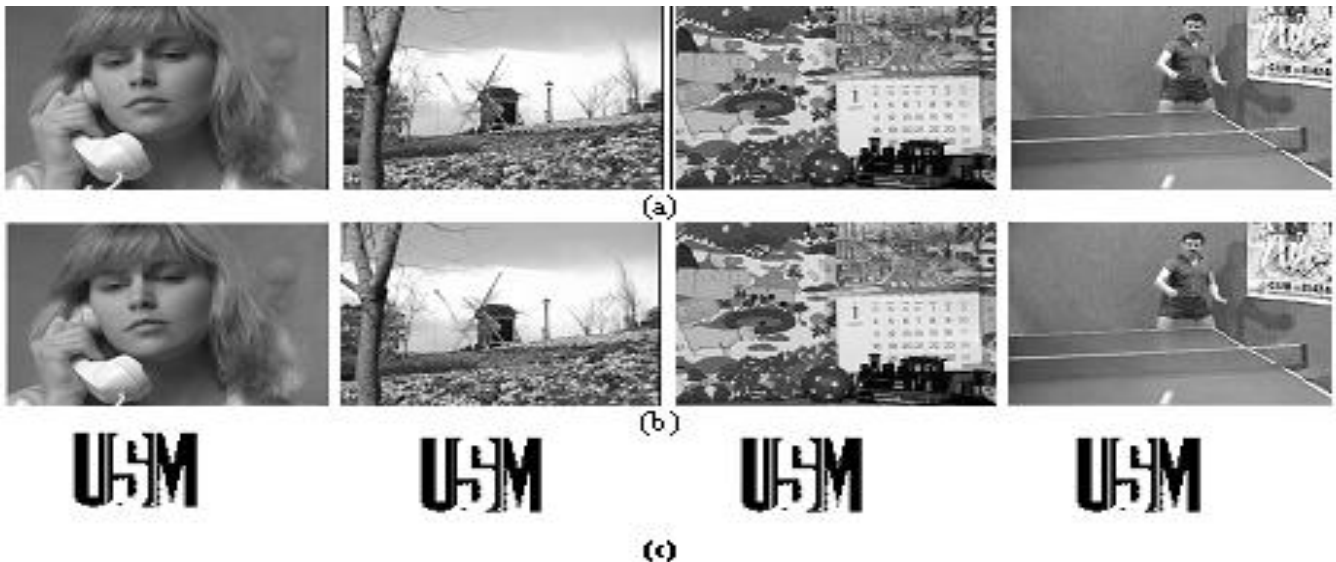
Figure 7. (a) original test frames of clips (b) watermarked test frames of clips (c) Extracted Watermark

From figure 7 can see no difference between the resolution of the original frame and watermarked frame.

Figure (8) shows the original watermark, extracted watermark without any threat with 1 correlation, and the detection score respectively.

(a) original watermark   (b) Extracted watermark

(c) Watermark detection results
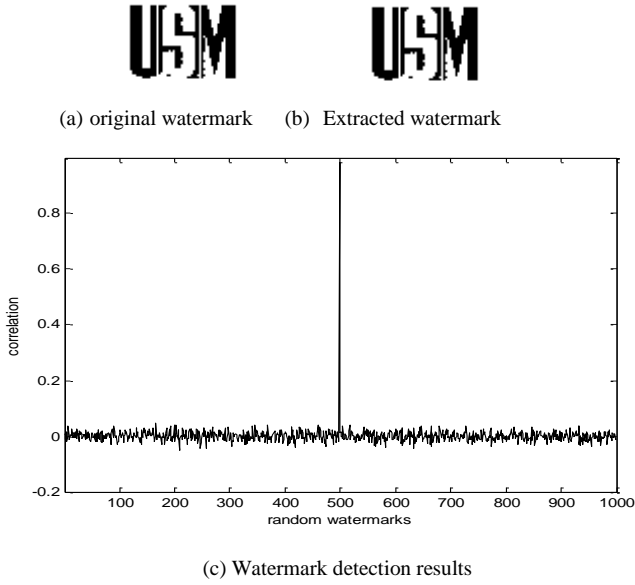
Figure 8. Watermark without any threat

A.   *Imperceptibility Results*

As a measurement for the quality of a watermarked frame, the peak signal to noise ratio (PSNR) is used. PSNR is defined as:

$$PSNR = 10\log(255^2 / MSE) \qquad (2)$$

$$MSE = \frac{1}{M.N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |X(i,j) - X^{\bullet}(i,j)|^2 \quad (3)$$

Where, X is the coefficients of the original video and X* are the coefficients of the watermarked video. M and N are the height and width of the frame respectively. In the proposed method, the watermark is embedded in the I-frame according to spatial domain. The average PSNR for all watermarked frame is 37.72dB. With this PSNR value, no quality degradation in the watermarked video is perceived.

B.   *Robustness Results*

Robustness is a measurement of the invulnerability of a watermark against the attempts to remove or degrade it by different types of geometric attacks. For the proposed method, the video watermarking application robustness is measured against geometric attacks, such as downscaling,

cropping, rotation, and frame dropping. Experimental results show that the proposed algorithm is very robust to geometric attacks; the similarity between the original and extracted watermarks is measured using the correlation factor *a* "NC", it may take between 0 and 1.

$$NC = \frac{\sum_{X} \sum_{Y} W.W^{*}}{\sum_{X} \sum_{Y} W.W} \qquad (4)$$

The similarity values vary in the interval [-1,1]; a value well above 0 and close to 1 indicates that the extracted sequence $W^{\bullet}$ matches the embedded sequence W. and therefore, we can conclude that the video has been watermarked with W.

1)   *Robust performance results against Downscaling*

The watermarked frame is scaled down to 50% with the aid of the bilinear interpolation method. Figures (9) show the watermarked frame, extracted watermark, and its detection score.

(a)Watermarked frame   (b) Extracted watermark

(c) Watermark detection results
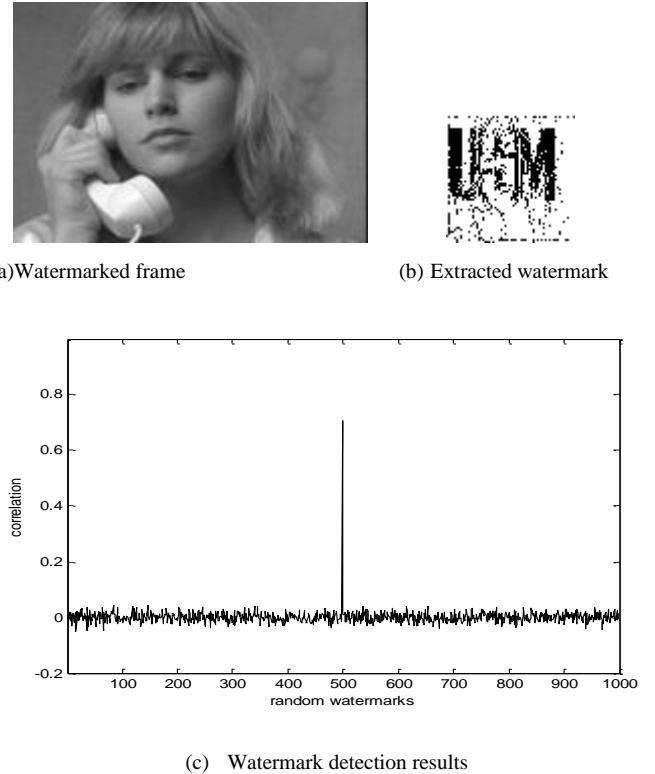
Figure 9. Watermarked frame under downscaling attack

2)   *Robust performance results against frame dropping*

The videos were segmented into seven scenes figure (10), assuming that seven watermarking groups were in need. The detection of the watermark after frame dropping of the extracted watermark is shown in the Figure (11) and the detection score has been shown in Figure (12).

Figure 10. Frames on the scene boundaries of the video: Susie, tennis, , flower, and mobile

Figure 11: Decoded watermark under frame dropping attack
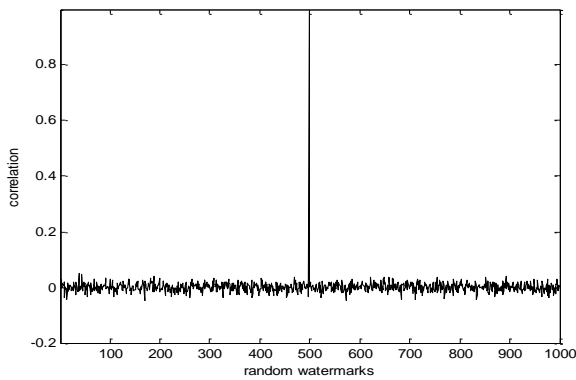
Figure 12 .Random watermark detection results under frame dropping attack

3) *Robust performance results against Cropping*

Cropping approximately 50% of the watermarked frame provides the covered watermark, although the correlation value is relatively small, the recovered logo can easily be distinguished, as shown in Figure (13).

(a) Watermarked frame   (b) Extracted watermark
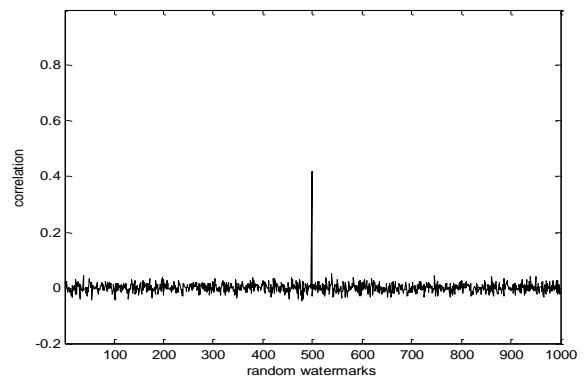
(c) Watermark detection results

Figure 13. Watermarked frame under cropping attack

4) *Robust performance results against Rotation.*

The watermark frame rotated by 5°,10°,15°,30° using bilinear interpolation extracted logo with correlation of

0.98, .97.97, 96 respectively, as shown in figure 14.

Figure 15 shows the watermarked frame rotated by −17° using bilinear interpolation, extracted logo with 0.99 correlation and detection score.

(a)Rotated counter clockwise with 5° (b) Rotated counter clockwise with 10°

(c)Rotated counter clockwise with 15° (d) Rotated counter clockwise with 30°

(e)Rotated clockwise with 5°   (f) Rotated clockwise with 10°

(g) Rotated clockwise with 15°   (h) Rotated clockwise with 30°

Figure 14. Watermarked frame under rotation attack

(a)Watermarked frame   (b) Extracted watermark
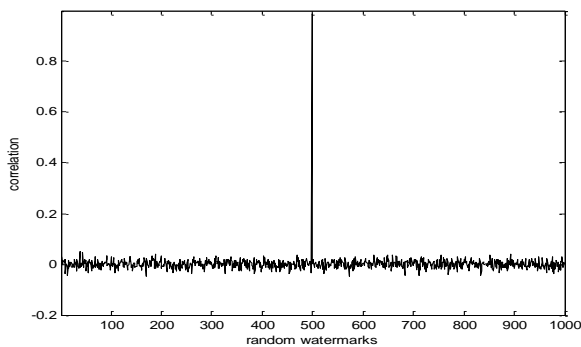
(c) Watermark detection results

Figure 15. Random watermark detection results under rotation attac

CONCLUSION

Robustness against geometric attacks is one of the most important requirements of the digital video watermark. In this paper, a novel robust video watermarking algorithm using spatial domain is proposed which embeds an image (logo) into host frames blocks, depending on the contrast of the block. The robustness of the proposed algorithm for video watermarking was illustrated against geometric attacks such as downscaling, cropping, rotation, and frame dropping. Simulation results demonstrated the effectiveness of the proposed method.

REFERENCES

[1] Harsh K Verma1, Abhishek Narain Singh2, Raman Kumar3 "Robustness of the Digital Image Watermarking Techniques against Brightness and Rotation Attack" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 5, No. 1, 2009.

[2] Voloshynovskiy S., Pereira S., Herrigel A., Baumgartner N., & Pun T. Generalized "watermarking attack based on watermark estimation and perceptual remodulation," SPIE 3971, Security and Watermarking of Multimedia Content II, San Jose, CA,2000.

[3] Voloshynovskiy S., Pereira S., Pun T., Eggers J J., & Su J. K. "Attacks on digital watermarks classification, estimation based attacks, and benchmarks," IEEE Communications Magazine,39(8),2001,pp. 118–126.

[4] Y Li,. X Gao,. Ji.H.,: "A 3D wavelet based spatial-temporal approach for video watermarking". in Proc. 5th Int. Conf. Computational Intelligence Multimedia Applications, Sep. 27–30, 2003, pp. 260–265

[5] Liu N .H.,. Chen J,. Huang X. Huang,. Shi .Y. Q,. :" A robust DWT based video watermarking algorithm". in Proc. IEEE Int. Symp. Circuits Systems, vol. 3, May 26–29, 2002, pp. 631–634.

[6] Hartung F., Girod B., "Watermarking of Uncompressed and Compressed Video" IEEE Transaction on Image Processing, Vol. 66, No. 3, 1998

[7] H. Inoue, A. Miyazaki, T. Araki, and T. Katsura, "A digital watermark method using the wavelet transform for video data," in Proc. IEEE Int. Symp. Circuits Systems, vol. 4, May 30, 1999, pp. 247–250.

[8] M. Ejima and A. Miyazaki, "A wavelet-based watermarking for digital images and video," in Proc. Int. Conf. Image Processing, vol. 3, Sep. 10–13, 2000, pp. 678–681.

[9] C. V. Ambroze, M. A. Tomlinson, and J. G. Wade, "Adding robustness to geometrical attacks to a wavelet based blind video watermarking system," in Proc. IEEE Int. Conf. Multimedia Expo, vol. 1, Aug. 26–29, 2002, pp. 557–560.

[10] C. V. Serdean, M. A. Ambroze, M. Tomlinson, and J. G. Wade, "DWT based high capacity blind video watermarking, invariant to geometrical attacks," in Proc. IEE Vision, Image, Signal Processing, vol. 150, 2003, pp. 51–58.

[11] Pik-Wah Chan and Michael R. Lyu1 ,"A DWT-based digital video watermarking scheme with error correction code". Fifth International Conference on Information and Communications Security ICICS 2003

[12] Patrizio Campisi and Alessandro Neri." perceptual video watermarking in the 3D-DWT domain using a multiplicative approach", IWDW 2005, LNCS 3710, pp. 432–443, 2005.

[13] Elbasi, E ,. Eskicioglu .A.M, "robust mpeg video watermarking in wavelet domain", Trakya Univ J Sci, 8(2): 87-93, 2007.

[14] Lv Anqiang, Li Jing, "A Novel Scheme for Robust Video Watermark in the 3D-DWT Domain", First International Symposium on Data, Privacy and E-Commerce, 2007.

[15] Xu Da-Wen, "A Blind Video Watermarking Algorithm Based on 3D Wavelet Transform", 2007 International Conference on Computational Intelligence and Security.

[16] Sadik. A.M .Al-Taweel ; Putra Sumari "Digital Video Watermarking in the Discrete Wavelet Transform Domain" Sixth International Conference on Computer Graphics, Imaging and Visualization, 2009, IEEE Computer Society. pp. 133–137, 2009.

[17] Sadik. Ali M.Al-Taweel, Putra Sumari, Saleh.Ali.K.Alomari, and Anas.J.A.Husain, "Digital Video Watermarking Algorithm Based on Discrete Cosine Transform Domain," Journal of Computer Science vol. 2,1,pp.23-28, 2009.

[18] Al-Taweel, S.A.M.; Sumari, P." Robust video watermarking based on 3D-DWT domain ",TENCON 2009 - 2009 IEEE Region 10 Conference

[19] A.Essaouabi, F.regragui, and E.Ibnelhaj, "A Wavelet-Based Digital Watermarking for Video," International Journal of Computer Science and Information Security (IJCSIS), vol. Vol. 6, No.1, 2009, 2009.

AUTHORS PROFILE

Sadik Ali M. Al-Taweel received the B.S. and M.S degree in Computer Sciences from Al-Mustansiriyah University and University of Technology in 1991 and 2003, respectively. During 2003-2005, he stayed at University of Science and Technology Yemen as an instructor of Computer Sciences and he worked as a lecturer. Currently he is a PhD student at the School of Computer Sciences, Universiti Sains Malaysia. He is a member of IEICE and IEEE reviewer of International Conference on Signal and Image Processing Applications (ICSIPA).

Putra Sumari obtained his MSc and PhD in 1997 and 2000 from Liverpool University, England. Currently, he is a lecturer at the School of Computer Science, Universiti Sains Malaysia, Penang. He is the head of the Multimedia Computing Research Group, School of Computer Science, USM. He is a member of ACM and IEEE, Program Committee and reviewer of several International Conference on Information and Communication Technology (ICT), Committee of Malaysian ISO Standard Working Group on Software Engineering Practice, Chairman of Industrial Training Program School of Computer Science USM, Advisor of Master in Multimedia Education Program, UPSI, Perak.
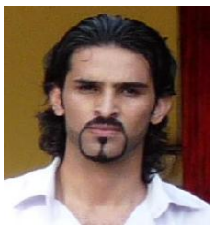
Saleh Ali K. Al-Omari Obtained his Bachelor degree in Computer Science from Jerash University, Jordan in 2004-2005 and Master degree in Computer Science from Universiti Sains Malaysia, Penang, Malaysia in 2007. Currently, He is a PhD candidate at the School of Computer Science, Universiti Sains Malaysia. His main research area interest now includes Peer to Peer Media Streaming, Video on Demand over Ad Hoc Networks, MANETs, and Multimedia Networking, Mobility.

An Energy Efficient Reliable Multipath Routing Protocol for Data Gathering In Wireless Sensor Networks

U.B. Mahadevaswamy

Assistant Professor, Department of Electronics and
communication
Sri Jayachamarajendra college of Engineering
Mysore, Karnataka, India.
mahadevaswamyphd@gmail.com
ubms_sjce@yahoo.co.in

M.N.Shanmukhaswamy

Professor
Department of Electronics and communication
Sri Jayachamarajendra college of Engineering,
Mysore, Karnataka, India.
mnsjce@gmail.com.

Abstract—**In Wireless Sensor Networks (WSN), the protocols that are accessible today have their own set of problems and most of them deal with energy efficiency. There is no specific work done on high network traffic or contention issue and significant work is remaining related to robustness and reliability. An important topic addressed by the wireless sensor networks community has been in-network data aggregation, because of the severe energy constraints of sensor nodes and the limited transport capacity of multihop wireless networks. In this paper, we propose to design an energy efficient reliable multipath routing protocol for data gathering in wireless sensor networks. This protocol is intended to provide a reliable transmission environment with low energy consumption, by efficiently utilizing the energy availability of the forwarding nodes to gather and distribute the data to sink, according to its requirements. By simulation results, we show that our proposed algorithm attains good packet delivery ratio with reduced energy consumption and delay**

Keywords- WSN, Multipath Routing Protocol, contention issue, sensor nodes, energy consumption, sink, Periodic Interest Propagation.

I. INTRODUCTION

Sensor nodes are those that have made the use of small, inexpensive, low-power, distributed devices, which are capable of local processing and wireless communication, a certainty in recent technological improvements [1]. Only a restricted amount of processing can be done by each sensor node. They have the ability to measure a given physical environment in vast detail, when coordinated with the information from a large number of other nodes. Thus, a sensor network can be defined as a collection of sensor nodes that co-ordinate to perform some specific task. The sensor networks depend on dense deployment and co- ordination to carry out their tasks in contrast to traditional networks. They have got a range of applications, which includes environmental monitoring – that involves monitoring air soil and water, condition based maintenance, habitat monitoring , seismic detection, military surveillance, inventory tracking, smart spaces etc. [1].

Due to recent advances in wireless sensor networks many new protocols were designed specifically for sensor networks for energy awareness. Since the routing protocol may vary depending on the application and network architecture, most of the attention has been given to them [2].

Data gathering is a common function of sensor networks in which the information is collected at sensor nodes and transported to central base stations for further processing and analysis. An important topic addressed by the wireless sensor networks community has been in-network data aggregation, because of the severe energy constraints of sensor nodes and the limited transport capacity of multihop wireless networks. To reduce expensive data transmission, sensor data has to be pre-processed in the network by the sensor nodes capable with computational power. Neglecting the characteristics of wireless transmission, most of the existing work on correlated data gathering completely assumes routing techniques similar to those in wire line networks [4].

The protocols that are accessible today have their own set of problems and most of them deal with energy efficiency. For high network traffic or contention issues, there is no work done. Significant work is remaining related to robustness and scalability. QoS routing have several applications including real time target tracking in battle environments, emergent event triggering in monitoring applications etc in sensor networks. At present, in an energy controlled environment like sensor networks, there is very little research that looks at managing QoS requirements.

To describe the class of routing mechanisms that let the establishment of multiple paths between source and destination, the term multipath routing has been used in the literature. For two reasons standard multipath routing has been explored. First the multi path routing is used in load balancing in which the traffic between a source-destination pair is split across multiple disjoint paths. Second use of multipath routing is to increase the chance of reliable data delivery. In these approaches, several copies of data are sent along diverse paths, to resist against failure of a certain number of paths [5].

In this paper, we propose to design an energy efficient reliable multipath routing algorithm for data gathering in wireless sensor networks. This protocol is intended to provide a reliable transmission environment with low energy consumption, by efficiently utilizing the energy availability of the forwarding nodes to gather and distribute the data to sink, according to its requirements.

II. RELATED WORK

Deepak Ganesan et al [5] have proposed a highly resilient; energy-efficient multipath routing protocol. It mainly discusses energy efficient recovery from failures, by discovering alternate paths. But it fails to consider the QoS qualities of the routes, when constructing multiple paths.

R Vidhyapriya et al [6], have developed an adaptive multipath routing protocol which spreads the traffic over the nodes lying on different possible paths between the source and the sink, in proportion to their residual energy and received signal strength. But it transmits all the packets across the multi paths, without considering the category of data.

Weifa Liang et al [7], have proposed a maximum network lifetime routing (MNL) algorithm to maximize the network lifetime while gathering online queries. In this protocol, the sink constructs a tree towards the source node, based on the residual energies of the nodes. But it does not consider the reliability of the transmitted data, since there will be large volumes of data involved, when there are continuous queries at the sink.

Ye Ming Lu et al [8], have proposed an energy efficient multipath routing algorithm which uses a load balancing algorithm to distribute the traffic over multiple disjoint paths. For energy efficiency, it uses the residual energy in the link cost function. But it does not consider aggregating similar data along multiple paths.

Yuzhe Liu et al [9], have proposed a priority based multipath routing protocol. It forwards the disseminated data based on the priority information accumulated hop count or remaining power resource. It uses either shortest path or energy-efficient path based on the priority tag. But this protocol does not consider the category of data and nature of queries.

Antoine B. Bagula et al [10], have proposed an energy constrained multipath routing. It minimizes the number of paths used in forwarding the data to the sink, there by minimizing the energy. But it does not discuss the sink's interest and reliability of data.

Octav Chipara et al [11], have proposed a real-time power aware routing (RPAR) protocol. It addresses important practical issues in wireless sensor networks, including lossy links, scalability, and severe memory and bandwidth constraints.

III. ENERGY EFFICIENT RELIABLE MULTIPATH ROUTING PROTOCOL

A. Protocol Overview

In this paper, we propose an energy efficient reliable multipath routing algorithm for data gathering in wireless sensor networks. It consists of three phases:

1. Periodic Interest Propagation by the sink
2. Energy Efficient Multipah Tree Construction
3. Packet Dispersion

In the first phase, the sink periodically broadcast an interest message containing its required data model, to all the nodes. In the second phase, we construct a multi path tree, in which nodes are selected based on their residual energy level. In the third phase, data sources of similar interests are gathered and transmitted towards the sink across the energy-efficient tree. When data sources cannot be aggregated, they are dispersed along multiple paths using erasure coding technique [11].

B. Periodic Interest Propagation by the Sink

A sink generates an interest message that identifies its requirement in wireless sensor networks which is then propagated throughout the network. On receiving an interest message, the source transmits the corresponding data. The data packets having similar interests are collected and aggregated at intermediate aggregators. The sink does not have any information on the availability of data while transmitting the first interest message. So the sink simply broadcasts interest message to all its neighbors. Interest message contains the Interest Id, Description and Timestamp. The features of shortest path algorithm can be used for interest message propagation.

An interest table is maintained by each node which contains the fields Interest Id, Sender Id, Cost of the message in terms of hop count and Timestamp. On receiving an interest message the node will look up in its interest for the received interest message. An interest table makes only one entry per data type from a particular sink.

When a node delivers the first interest message, it is added in the interest table with its parameters. The interest message is then rebroadcast to other nodes. It checks the interest table, if a duplicate interest message is received by a node. The duplicate message is dropped when the cost of it is higher than the cost of the earlier message; else it is updated in the table and then forwarded to the next node.

The proposed protocol consists of a periodic interest propagation phase. Since the interest is a soft state, it is very often refreshed by the sink. Refreshing is essential since it is impossible to transmit interest reliably across the network and the refresh rate is a protocol design parameter. To propagate the interest based on the previously cached data, either flooding or directional propagation may be used.

C. Energy Efficient Multipath Tree Construction

We propose a heuristic algorithm for the tree construction. We consider the wireless sensor network M as a directed graph G (N, E). Let the set of nodes N consisting of sensors

and (a, b) $\in$ E if a and b are residing inside the transmission range of each other. The fundamental idea of the proposed algorithm is, when a data gathering request is arrived, then using the greedy algorithm a data gathering tree for the request is constructed. The greedy algorithm maximizes the minimum residual energy among the nodes. Then the nodes are included in the tree one by one but in beginning only the sink node is included. A node b is selected to be included into the tree if causes to maximize the minimum residual energy among the trees including it.

In our algorithm, we use the following notations

- N is the total number of nodes
- N_T is the set of nodes in the tree,
- stop is a Boolean variable,
- newnode is the node that will be added to the tree.
- q is the size of the sensed data by newnode.
- $w^\alpha_{a,b}$ is the weight assigned to the edge.
- R is the set of nodes that are not in the tree.
- RE is the residual energy.
- s is the sink node
- mre_{max} is the maximum value of minimum residual energy at each node of the tree.
- tp is the temporary parent node.
- $P_{a,s}$ is the unique path in T from node a to node s
- p(a) is the parent of a in T
- Let node v $\in$ N - N_T be the considered node.

1) Tree Construction Algorithm

Algorithm: 1

1. $N_T = \{s\}$

2. *stop* =" *false* "

3. $R = N - N_T$

4. $RE(s) = \infty$

5. $mre_{max} = 0$

6. *For* each i $\in$ R

6.1 *Compute* $mre_{max}(i)$ and tp

 6.2. *If* $mre_{max}(i) > mre_{max}$, then

  6.2.1. $mre_{max} = mre_{max}(i)$

  6.2.2. $Newnode = i$

 6.3 *End if*

7. *End* for

8 *If* $mre_{max} > 0$, then

 8.1. $P(newnode) = tp(newnode)$

 8.2. *For* each j $\in$ $P_{newnode, s}$ *do*

  8.2.1. $RE(j) = RE(j) - qw^\alpha_{j,p(j)}$

 8.3 *End* for

 8.4. $N_T = N_T \cup \{newnode\}$

 8.5 $R = R - newnode$

9 *Else*

 9.1 *stop* = "*True*"

10. *End if*

11. *If* $(R \neq \phi)$ or stop = "false" then

 11.1 *repeat* from 5

12 *End if*

13 End

D. Packet Dispersion

The simplified message manipulation and the reliable data transmission are the advantages of using the dispersion algorithm [12] and erasure code [13].

We propose a new packet dispersion mechanism which splits the data packets at the source into fragments and distributes them on the multiple parallel paths, in order to reduce the packet loss. The packets are reassembled at the destination. Based on robin dispersal algorithm, we have to utilize an erasure code technique in order to make this mechanism efficient [14].

The source node breaks up the packet into N fragments of size s, generates K fragments of parity and transmits the total of N+K packets to the destination. The destination must receive at least N fragments within T_m time units in order to make the transmission to be successful.

Through the stronger paths the important fragments can be sent between the replicated fragments. If any unexpected fault takes place then the appropriate stronger paths can be chosen from the list.

IV. OVERALL ALGORITHM

The following algorithm summarizes the overall process of our proposed approach.

Let n1, n2.... be the *N* sensor nodes

Let $d(n1, n2)$ be the distance between the nodes *n*1 and *n*2 .

Algorithm: 2

1. Sink periodically broadcasts the interest message.
2. Nodes receive the interest message.
3. It checks whether it is present already in its table.
 3.1 If not exist, then
  3.1.1 Add into its table
 3.2 Else
  3.2.1 Rebroadcast to its neighbors.
 3.3 End if
4. Suppose if a query arrives, an energy efficient tree is constructed using algorithm 1.
5. Each node checks its interest table which matches the query.
6. The matched data is sent to its downstream nodes.
7. If $d(n_i, n_j) < D$, where n_i and n_j are two sensors with matching data, then
 . 7.1 The matched data is gathered from all the corresponding nodes and sent to the
  sink via the tree.
8. Else

8.1 The data is dispersed and transmitted along multiple paths to the sink.
9. End if

V. SIMULATION RESULTS

A. Simulation Setup

The performance of EERMR protocol is evaluated through NS2 simulation. A random network deployed in an area of 500 X 500 m is considered. We vary the number of nodes as 20, 40….100. Initially the nodes are placed randomly in the specified area. The base station is assumed to be situated 100 meters away from the above specified area. The initial energy of all the nodes assumed as 3.1 joules. In the simulation, the channel capacity of mobile hosts is set to the same value: 2 Mbps. The distributed coordination function (DCF) of IEEE 802.11 is used for wireless LANs as the MAC layer protocol. The simulated traffic is CBR with UDP source and sink. The number of sources is varied from 1 to 4.

Table 1 summarizes the simulation parameters used

TABLE I: SIMULATION PARAMETERS

No. of Nodes	20,40,….100
Area Size	500 X 500
Mac	802.11
Simulation Time	50 sec
Traffic Source	CBR
Packet Size	512
Transmit Power	0.660 w
Receiving Power	0.395 w
Idle Power	0.335 w
Initial Enegy	3.1 J
Transmission Range	75m

B. Performance Metrics

The performance of ERRMR is compared with the MNL and SPT [7] protocols. The performance is evaluated mainly, according to the following metrics.

Control Overhead: The control overhead is defined as the total number of routing control packets normalized by the total number of received data packets.

Average end-to-end Delay: The end-to-end-delay is averaged over all surviving data packets from the sources to the destinations.

Average Packet Delivery Ratio: It is the ratio of the number .of packets received successfully and the total number of packets transmitted.

Energy Consumption: It is the average energy consumption of all nodes in sending, receiving and forward operations

The simulation results are presented in the next section.

C. Simulation Results

A. Based on Nodes

In our initial experiment, we vary the number of nodes as 20, 40, 60, 80 and 100.
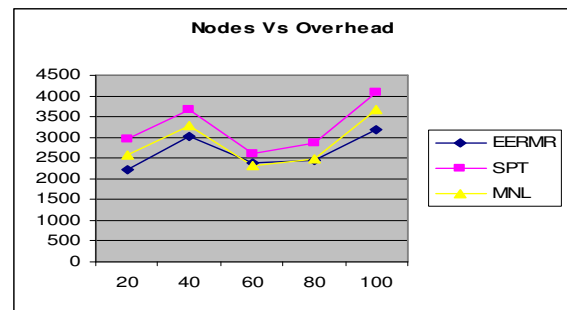
Figure 1. Nodes Vs Overhead
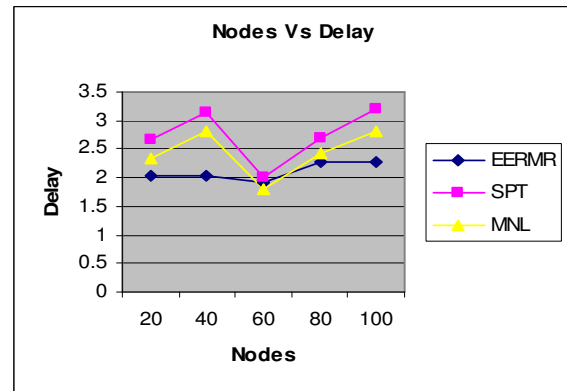
Figure 2. Nodes Vs Delay
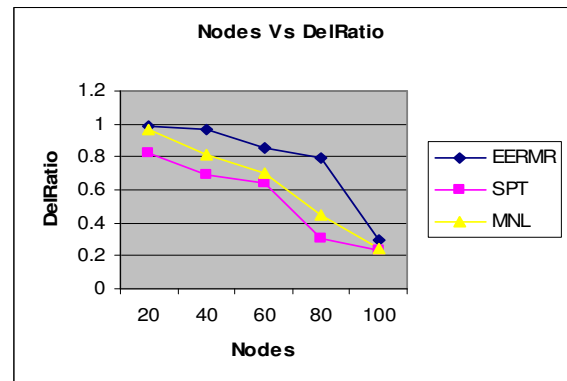
Figure 3. Nodes Vs DelRatio
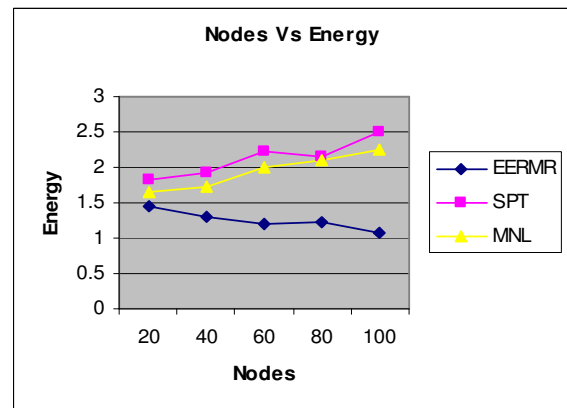
Figure 4. Nodes Vs Energy

Figure 1 gives the control overhead occurred for all the protocols when the number of nodes are increased. From the figure, we can ensure that the control overhead is less for EERMR when compared to other protocols.

Figure 2 gives the average end-to-end delay for all the protocols when the number of nodes is increased. From the figure, it can be seen that the average end-to-end delay of the proposed EERMR protocol is less when compared with all other protocols.

Figure 3 presents the packet delivery ratio of all the protocols. Since reliability is achieved using the dispersion technique, EERMR achieves good delivery ratio, compared to other protocols.

Figure 4 shows the results of energy consumption for all the protocols. From the results, we can see that EERMR protocol has less energy consumption than all other protocols, since it has the energy efficient tree.

B. Based on Sources

In the second experiment, we vary the number of sources as 1, 2, 3, and 4.
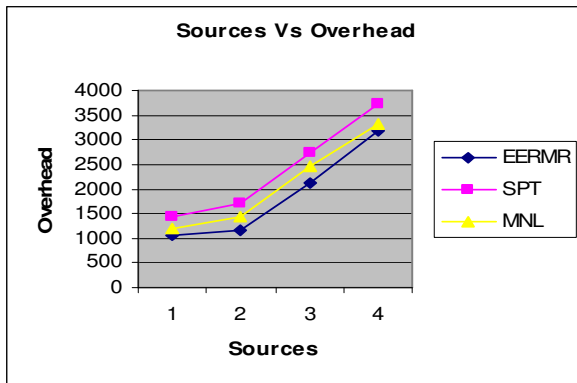
Figure 5. Sources Vs Overhead
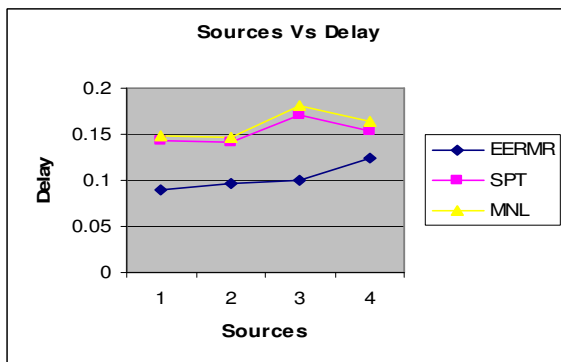
Figure 6. Sources Vs Delay
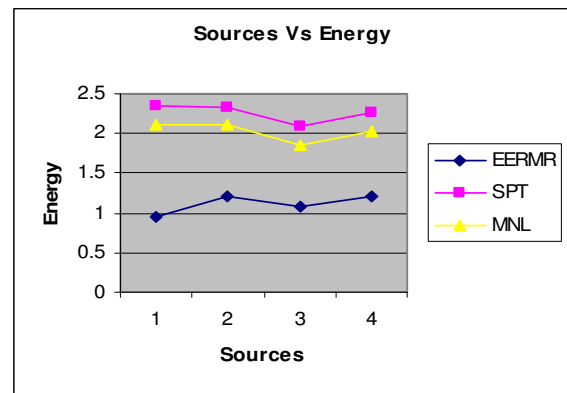
Figure 7. Sources Vs Energy
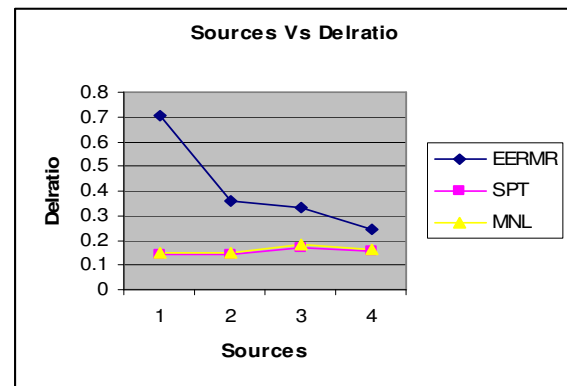
Figure 8: Sources Vs Del Ratio

From Figure 5, we can ensure that the control overhead is less for EERMR when compared with other protocols.

From Figure 6, we can see that the average end-to-end delay of the proposed EERMR protocol is less when compared with all other protocols.

Figure 7 shows the results of energy consumption for all the protocols. From the results, we can see that EERMR protocol has less energy consumption than all other protocols, since it has the energy efficient routing.

Figure 8 gives the packet delivery ratio of all the protocols. Since reliability is achieved using the dispersion technique, EERMR achieves good delivery ratio, compared to other protocols when the number of sources are increased.

VI. CONCLUSION

In this paper, we have proposed an energy efficient reliable multipath routing protocol for data gathering in wireless sensor networks. The proposed protocol provides a reliable transmission environment with low energy consumption, by efficiently utilizing the energy availability of the forwarding nodes to gather and distribute the data to sink, according to its requirements. In this approach, the sink periodically broadcast an interest message containing its required data model, to all the nodes. A multi path tree is constructed in which nodes are selected based on their residual energy level. Then data sources of similar interests are gathered and transmitted across the energy-efficient tree towards the sink. When data sources

cannot be aggregated, they are dispersed along multiple paths using erasure coding technique. By simulation results, we have shown that our proposed algorithm attains good packet delivery ratio with reduced energy consumption and delay. In our future work, we wish to apply some compression techniques in data gathering to reduce the delay. Also we shall use some trusting mechanism such that the accuracy of gathered data is increased

REFERENCES

[1] Archana Bharathidasan and Vijay Anand Sai Ponduru, "Sensor Networks: An Overview", IEEE INFOCOM 2004.

[2] Kemal Akkaya and Mohamed Younis, "A Survey on Routing Protocols for Wireless Sensor Networks", Elseiver 2003.

[3] Chalermek Intanagonwiwat, Ramesh Govindan, Deborah Estrin, "Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks", In Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking (MOBICOM '00), August 2000, Boston, Massachussetts.

[4] Tao Cui, Lijun Chen, Tracey Ho, Steven H. Low, and Lachlan L. H. Andrews, "Opportunistic Source Coding for Data Gathering in Wireless Sensor Networks", Proceedings of the 2007 IEEE conference on Diversity in computing, IEEE 2007.

[5] Deepak Ganesan, Ramesh Govindan, Scott Shenker and Deborah Estrin, "Highly-Resilient Energy- Efficient Multipath Routing in Wireless Sensor Networks", Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing., Volume 5 , Issue 4 ,October 2001

[6] R Vidhyapriya and Dr P T Vanathi, "Energy Efficient Adaptive Multipath Routing for Wireless Sensor Networks", IAENG International Journal of Computer Science, 15 August 2007.

[7] Weifa Liang and Yuzhen Liu, "Online Data Gathering for Maximizing Network Lifetime in Sensor Networks", IEEE Transactions on Mobile Computing, January- 2007

[8] Ye Ming Lu and Vincent W.S. Wong, "An Energy-Efficient Multipath Routing Protocol for Wireless Sensor Networks", International Journal of Communication Systems, July-2007.

[9] Yuzhe Liu and Yuzhe Liu, "A Priority-based Multi-path Routing Protocol for Sensor   Networks", IEEE 15th International Symposium on Personal, Indoor and Mobile Radio Communications, September-2004

[10] Antoine B. Bagula and Kuzamunu G. Mazandu, "Energy Constrained Multipath   Routing in Wireless Sensor Networks", Springer-Verlag Berlin Heidelberg, 2008

[11] Octav Chipara, Zhimin He, Guoliang Xing, Qin Chen and Xiaorui Wang, "Real-time Power-Aware Routing in Sensor Networks", 14th IEEE International Workshop on Quality of Service, IWQoS 2006, June-2006

[12] Panagiotis Papadimitratos and Zygmunt J. Haas, "Secure Data Communication in Mobile Ad Hoc Networks", IEEE Journal On Selected Areas In Communications, Vol. 24, No. 2, February 2006.

[13] Panagiotis Papadimitratos and Zygmunt J. Haas, "Secure message transmission in mobile ad hoc networks ", 2003 Elsevier, Ad Hoc Networks 1 (2003) 193–209.

[14] M.O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance", J. ACM 36 (2) (1989) 335–348

Mr.U.B.Mahadevaswamy completed his B.E. degree in Electronics and Communication from Mysore University in the year 1988, M.Tech in Industrial Electronics from Manglore University in the year 1995 and He is presently working as Assistant Professor in the Department of Electronics and communication, Sri Jayachamarajendra college of Engineering, Mysore, Karnataka, India. He is doing his Ph. D in the area of Wireless sensor networks under the guidance of Dr. M.N Shanmukhaswamy. His field of interest includes Wireless sensor networks, Analog and mixed mode VLSI circuits, Control systems, Digital signal processing.

Dr.M.N.Shanmukha Swamy completed his B.E. degree in Electronics and Communication from Mysore University in the year 1978, M.Tech in Industrial Electronics from the same university in the year 1987 and obtained his Ph.D in the field of Composite materials from Indian Institute of Science, Bangalore in 1997. He is presently working as Professor in the Department of Electronics and communication, Sri Jayachamarajendra college of Engineering, Mysore, Karnataka, India. He is guiding several research scholars and has published many books & papers both in National & International conferences & journals. His research area includes Wireless Sensor Networks, Biometrics, VLSI and composite materials for application in electronics.

A Novel Approach towards Cost Effective Region-Based Group Key Agreement Protocol for Secure Group Communication

K. Kumar
Research Scholar &
Lecturer in CSE
Government College of Engg,
Bargur- 635104, Tamil Nadu,
India
pkk_kumar@yahoo.com

J. Nafeesa Begum
Research Scholar &
Sr. Lecturer in CSE
Government College of Engg,
Bargur- 635104, Tamil Nadu,
India
nafeesa_jeddy@yahoo.com

Dr.V. Sumathy
Asst .Professor in ECE
Government College of
Technology,
Coimbatore, Tamil Nadu,
India
sumi_gct2001@yahoo.co.in

Abstract—**This paper addresses an interesting security problem in wireless ad hoc networks: the Dynamic Group Key Agreement key establishment. For secure group communication in an Ad hoc network, a group key shared by all group members is required. This group key should be updated when there are membership changes (when the new member joins or current member leaves) in the group. In this paper, We propose a novel, secure, scalable and efficient Region-Based Group Key Agreement protocol (RBGKA) for ad-hoc networks. This is implemented by a two-level structure and a new scheme of group key update. The idea is to divide the group into subgroups, each maintaining its subgroup keys using Group Diffie-Hellman (GDH) Protocol and links with other subgroups in a Tree structure using Tree-based Group Diffie-Hellman (TGDH) protocol. By introducing region-based approach, messages and key updates will be limited within subgroup and outer group; hence computation load is distributed among many hosts. Both theoretical analysis and experimental results show that this Region-based key agreement protocol performs better for the key establishment problem in ad –hoc network in terms of memory cost, computation cost and communication cost.**

Keywords- Ad Hoc Network, Region-Based Group Key Agreement Protocol, Group Diffie-Hellman, Tree-Based Group Diffie-Hellman.

I. INTRODUCTION

Wireless networks are growing rapidly in recent years. Wireless technology is gaining more and more attention from both academia and industry. Most wireless networks used today e.g the cell phone networks and the 802.11 wireless LAN, are based on the wireless network model with pre-existing wired network infrastructures. Packets from source wireless hosts are received by nearby base stations, then injected into the underlying network infrastructure and then finally transferred to destination hosts.

Another wireless network model, which is in active research, is the ad-hoc network. This network is formed only by mobile hosts and requires no pre-existing network infrastructure. Hosts with wireless capability form an ad- hoc network, some mobile hosts work as routers to relay packets from source to destination. It is very easy and economic to form an ad-hoc network in real time. Ad-hoc network is ideal in situations like battlefield or rescuer area where fixed network infrastructure is very hard to deploy.

A mobile ad hoc network is a collection of autonomous nodes that communicate with each other. Mobile nodes come together to form an ad hoc group for secure communication purpose. A key distribution system requires a trusted third party that acts as a mediator between nodes of the network. Ad-hoc networks characteristically do not have a trusted authority. Group Key Agreement means that multiple parties want to create a common secret key to be used to exchange information securely. Furthermore, group key agreement also needs to address the security issue related to membership changes due to node mobility. The membership change requires frequent changes of group key. This can be done either periodically or updating every membership changes. The changed group key ensures backward and forward secrecy. With frequent changes in group memberships, the recent researches began to pay more attention on the efficiency of group key update. Recently, collaborative and group –oriented applicative situations like battlefield, conference room or rescuer area in mobile ad hoc networks have been a current research area. Group key agreement is a building block in secure group communication in ad hoc networks. However, group key agreement for large and dynamic groups in ad hoc networks is a difficult problem because of the requirements of scalability and security under constraints of node available resources and node mobility.

We propose a communication and computation efficient group key agreement protocol in ad-hoc network. In large and high mobility ad hoc networks, it is not possible to use a single group key for the entire network because of the enormous cost of computation and communication in rekeying. So, we divide the group into several subgroups; let each subgroup has its subgroup key shared by all members of the subgroup. Each group has sub group controller node and gateway node, in

which the sub group controller node is controller of subgroup and gateway node is controller among subgroups. Let each gateway member contribute a partial key to agree with a common Outer group key among the subgroups.

The contribution of this work includes:

1. In this paper, we propose a new efficient method for solving the group key management problem in ad-hoc network. This protocol provides efficient, scalable and reliable key agreement service and is well adaptive to the mobile environment of ad-hoc network.

2. We introduce the idea of subgroup and subgroup key and we uniquely link all the subgroups into a tree structure to form an outer group and outer group key. This design eliminates the centralized key server. Instead, all hosts work in a peer-to-peer fashion to agree on a group key. We use Region-Based Group Key Agreement (RBGKA) as the name of our protocol. Here we propose a region based group key agreement protocol for ad hoc networks called Region-Based GDH & TGDH protocol.

3. We design and implement Region-Based Group key agreement protocol using Java and conduct extensive experiments and theoretical analysis to evaluate the performance like memory cost, communication cost and computation cost of our protocol for Ad- Hoc network.

The rest of the paper is as follows, Section II briefly presents various group key agreement protocols. Section III presents the proposed schemes. Section IV describes the Experimental Results and Discussion. Section V describes the Performance analysis and finally Section VI concludes the paper.

II. RELATED WORK

Steiner et al. [1,2,3] proposed CLIQUES protocol suite that consist of group key agreement protocols for dynamic groups called Group Diffie-Hellman(GDH). It consists of three protocols namely GDH.1, GDH.2 and GDH.3. These protocols are similar since they achieve the same group key but the difference arises out of the computation and communication costs. Yongdae Kim et al. [4, 8] proposed Tree-Based Group Diffie-Hellman (TGDH) protocol, wherein each member maintains a set of keys arranged in a hierarchical binary tree. TGDH is scalable and require a few rounds (O (log (n)) for key computation but their major drawback is that they require a group structure and member serialization for group formation. Ingemarsson et al in [5] proposed the protocol referred to as ING. This Protocol executes in n-1 rounds and requires the members to be arranged in a logical ring. The advantages of this scheme are that there is no Group Controller, every member does equal work and the message size is constant. On the other hand, the protocol suffers from communication overhead, inefficient join/leave operations and the requirements for a group structure which is difficult to realize in Ad hoc networks. Another protocol for key agreement was proposed in [6] by Burmester and Desmedt. The protocol involves two broadcast rounds before the

members agree on a group key. This scheme has several advantages such as the absence of a GC, equal work load for key establishment and a small constant message size. Some of the drawbacks of this scheme are that it requires the member to be serialized, different workload for join/leave and it is not very efficient. The Skinny Tree (STR) protocol proposed by steer et al. in [7] and undertaken by Kim et al. in [8], is a Contributory protocol. The leave cost for STR protocol is computed on average, since it depends on the depth of the lowest numbered leaving member node.

The group key agreement protocols provide a good solution to the problem of managing keys in Ad hoc networks as they provide the ability to generate group key which adapts well to the dynamic nature of ad hoc network groups. The group key agreement is not so easy to implement in ad hoc network environments because it has some special characteristics that these networks have. Thus one has to meet the security goals and at the same time should not fail to remember the computational and communication limitations of the devices. Regarding the Group Key Agreement protocols, it is easy to note that one single protocol cannot meet the best of the needs of all kinds of ad hoc networks.

In this paper, we propose a combination of two protocols that are well suited to ad hoc networks [9]. This paper uses the GDH.2 and TGDH protocols. The GDH.2 protocols are attractive because these do not involve simultaneous broadcast and round synchronization. The costs in TGDH are moderate, when the key tree is fully balanced. Therefore, these are well suited for dynamic membership events in ad hoc networks.

III. PROPOSED SCHEME

A. Motivation

There has been a growing demand in the past few years for security in collaborative environments deployed for emergency services where our approach can be carried out very efficiently is shown in Fig.1.Confidentiality becomes one of the top concerns to protect group communication data against passive and active adversaries. To satisfy this requirement, a common and efficient solution is to deploy a group key shared by all group application participants. Whenever a member leaves or joins the group, or whenever a node failure or restoration occurs, the group key should be updated to provide forward and backward secrecy. Therefore, a key management protocol that computes the group key and forwards the rekeying messages to all legitimate group members is central to the security of the group application.

Figure.1. Secure Group Applications

In many secure group applications, a Region based contributory GKA schemes may be required. In such cases,

the group key management should be both efficient and fault-tolerant. In this paper, we describe a military scenario (Figure.2). A collection of wireless mobile devices are carried by soldiers or Battlefield tanks. These mobile devices cooperate in relaying packets to dynamically establish routes among themselves to form their own network "on the fly". However, all nodes except the one with the tank, have limited battery power and processing capacities. For the sake of power- consumption and computational efficiency, the tank can work as the Gateway member while a contributed group key management scheme is deployed.

Figure.2. Battlefield Scenario

B. System Model

a) Overview of Region-Based Group Key Agreement Protocol:

The goal of this paper is to propose a communication and computation efficient group key establishment protocol in ad-hoc network. The idea is to divide the multicast group into several subgroups, let each subgroup has its subgroup key shared by all members of the subgroup. Each Subgroup has subgroup controller node and a Gateway node, in which Subgroup controller node is the controller of subgroup and a Gateway node is controller of subgroups controller.

For example, in Figure.3, all member nodes are divided into number of subgroups and all subgroups are linked in a tree structure as shown in Figure.4.
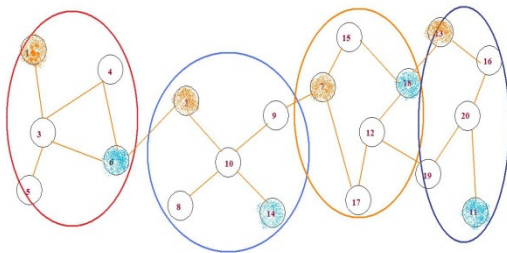
Figure.3: Members of group are divided into subgroups
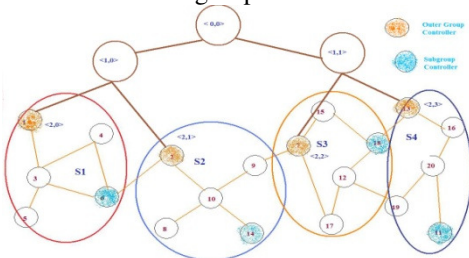
Figure.4: Subgroups link in a Tree Structure

The layout of the network is as shown in below figure.5.
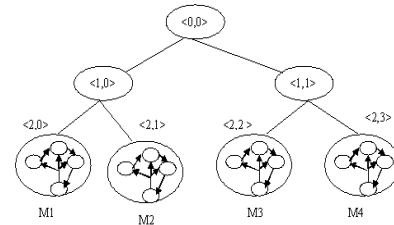
Figure.5. Region based Group Key Agreement

One of the members in the subgroup is subgroup controller. The last member joining the group acts as a subgroup controller. Each outer group is headed by the outer group controller. In each group, the member with high processing power, memory, and Battery power acts as a gateway member. Outer Group messages are broadcast through the outer group and secured by the outer group key while subgroup messages are broadcast within the subgroup and secured by subgroup key.

Let N be the total number of group members, and M be the number of the subgroups in each subgroup, then there will be N/M subgroups, assuming that each subgroup has the same number of members.

There are two shared keys in the Region-Based Group Key Agreement Scheme:

1. Outer Group Key (KG)is used to encrypt and decrypt the messages broadcast among the subgroup controllers.
2. The Subgroup Key (KR) is used to encrypt and decrypt the Sub Group level messages broadcast to all sub group members.

In our Region-Based Key Agreement protocol shown in Fig.5 a Subgroup Controller communicates with the member in the same region using a Regional key (i.e Sub group key) KR. The Outer Group key KG is derived from the Outer Group Controller. The Outer Group Key KG is used for secure data communication among subgroup members. These two keys are rekeyed for secure group communications depending on events that occur in the system.

Assume that there are totally N members in Secure Group Communication. After sub grouping process (Algorithm 1), there are S subgroups $M_1, M_2 \ldots M_s$ with $n_1, n_2 \ldots n_s$ members.

Algorithm. 1. Region-Based Key Agreement protocol

1. The Subgroup Formation
    The number of members in each subgroup is
    N / S < 100.
Where,
N – is the group size. and
   S – is the number of subgroups.
 Assuming that each subgroup has the same number of members.
2. The Contributory Key Agreement protocol is implemented among the group members. It consists of three stages.
    a. To find the Subgroup Controller for each subgroups.

---

b. GDH protocol is used to generate one common key for each subgroup headed by the subgroup controller.

c. Each subgroup gateway member contributes partial keys to generate a one common backbone key (i.e Outer group Key (KG)) headed by the Outer Group Controller using TGDH protocol.

3. Each Group Controller (Sub /Outer) distributes the computed public key to all of its members. Each member performs rekeying to get the corresponding group key.

A Regional key KR is used for communication between a subgroup controller and the members in the same region. The Regional key KR is rekeyed every time whenever there is a membership change event, subgroup join / leave and member failure. The Outer Group key KG is rekeyed whenever there is a join / leave subgroup controllers and member failure to preserve secrecy.

The members within a subgroup use Group Diffie-Hellman Contributory Key Agreement (GDH). Each member within a subgroup contributes his share in arriving at the subgroup key. Whenever membership changes occur, the subgroup controller or previous member initiates the rekeying operation.

The gateway member initiates communication with the neighboring members belonging to another subgroup and mutually agree on a key using Tree-Based Group Diffie-Hellman contributory Key Agreement(TGDH) protocol to be used for inter subgroup communication between the two subgroups. Any member belonging to one subgroup can communicate with any other member in another subgroup through this member as the intermediary. In this way adjacent subgroups agree on outer group key. Whenever membership changes occur, the outer group controller or previous group controller initiates the rekeying operation.

Here, we prefer the subgroup key to be different from the key for backbone. This difference adds more freedom of managing the dynamic group membership. Additionally, by using this approach one can potentially save the communication and computational cost.

C .Network Dynamics

The network is dynamic in nature. Many members may join or leave the group. In such cases, a group key management system should ensure that backward and forward secrecy is preserved.

1. Member Join

When a new member joins, it initiates communication with the subgroup controller. After initialization, the subgroup controller changes its contribution and sends public key to this new member. The new member receives the public key and acts as a group controller by initiating the rekeying operations for generating a new key for the subgroup. The rekeying operation is as follows.

$$\text{New node} \xrightarrow{\text{Join request}} \text{Subgroup Controller}$$

$$\textit{Subgroup Controller} \xrightarrow{\text{change its contribution and send public key to}} \text{New Node}$$

$$\text{New Node} \xrightarrow{\text{Acts as}} \text{New Subgroup Controller}$$

$$\text{New Subgroup Controller} \xrightarrow[\text{Multicast this public key value to}]{\text{puts its contribution to all the public key value \&}} \text{the entire member in the subgroup}$$

$$\text{Each Member} \xrightarrow{\text{put is contribution to the public value \& Compute}} \text{New Subgroup Key}$$

2.Member Leave:

a)When a Subgroup member Leaves

When a member leaves subgroup to which it belongs the subgroup key must be changed to preserve the forward secrecy. The leaving member informs the subgroup controller. The subgroup controller changes its private key value, computes the public value and broadcasts the public value to all the remaining members. Each member performs rekeying by putting its contribution to public value and computes the new Subgroup Key. The rekeying operation is as follows.

$$\text{Leaving Node} \xrightarrow{\text{Leaving Message}} \text{Subgroup Controller}$$

$$\text{Subgroup Controller} \xrightarrow[\text{Multicast the public key value to}]{\text{changes its private key value, compute the public key value and}} \text{All the remaining Member}$$

$$\text{Each Member} \xrightarrow{\text{Performs Rekeying and Compute}} \text{New Subgroup Key}$$

b)When Subgroup Controller Leaves:

When the Subgroup Controller leaves, the Subgroup key used for communication among the subgroup controllers needs to be changed. This Subgroup Controller informs the previous Subgroup Controller about its desire to leave the subgroup which initiates the rekeying procedure. The previous subgroup controller now acts as a Subgroup controller. This Subgroup controller changes its private contribution value and computes all the public key values and broadcasts to all the remaining members of the group. All subgroup members perform the rekeying operation and compute the new subgroup key. The rekeying operation is as follows.

$$\text{Leaving Subgroup Controller} \xrightarrow{\text{Leaving Message}} \text{Old Subgroup Controller}$$

$$\text{Old Subgroup Controller} \xrightarrow[\text{public key value and Multicast}]{\text{change its private value,compute the all}} \text{Remaining Member in the group}$$

$$\text{Subgroup Member} \xrightarrow{\text{Perform Rekeying and Compute}} \text{New Subgroup Key}$$

c) When Outer Group Controller Leaves:

When a Outer group Controller leaves, the Outer group key used for communication among the Outer groups needs to be changed. This Outer group Controller informs the previous Outer group Controller about its desire to leave the Outer group which initiates the rekeying procedure. The previous Outer Group controller now becomes the New Outer group controller. This Outer group controller changes its private contribution value and computes the public key value and broadcast to the entire remaining member in the group. All Outer group members perform the rekeying operation and compute the new Outer group key. The rekeying operation is as follows.

$$\text{Leaving Outer group Controller} \xrightarrow{\text{Leaving Message}} \text{Old Outer group Controller}$$

Old Outer group Controller $\xrightarrow{\text{change its private value,compute the all public key value and Multicast}}$ Remaing Member in the Outer group

Outer group Member $\xrightarrow{\text{Perform Rekeying and Compute}}$ New Outer group Key

d) When Gateway member leaves

When a gateway member leaves the subgroup, it delegates the role of the gateway to the adjacent member having high processing power, memory, and Battery power and the adjacent member acts as a new gateway member. Whenever the gateway member leaves, all the two keys should be changed. These are

i. Outer group key among the subgroups.
ii. Subgroup key within the subgroup.

In this case, the subgroup controller and outer group controller perform the rekeying operation. Both the Controller leave the member and a new gateway member is selected in the subgroup, performs rekeying in the subgroup. After that, it joins in the outer group. The procedure is same as member join in the outer group.

D. Communication Protocol:

The members within the subgroup have communication using subgroup key. The communication among the subgroup members takes place through the gateway member.

1. Communication within the Subgroup:

The sender member encrypts the message with the subgroup key (KR) and multicasts it to all members in the subgroup. The subgroup members receive the encrypted message, perform the decryption using the subgroup key (KR) and get the original message. The communication operation is as follows.

Source Member $\xrightarrow{E_{KR}[\text{Message}] \ \& \ \text{Multicast}}$ Destination Member

Destination Member $\xrightarrow{D_{KR}[E_{KR}[\text{Message}]]}$ Original Message

2. Communication among the Subgroup:

The sender member encrypts the message with the subgroup key (KR) and multicasts it to all members in the subgroup. One of the members in the subgroup acts as a gate way member. This gateway member decrypts the message with subgroup key and encrypts with the outer group key (KG) and multicasts to the entire gateway member among the subgroup. The destination gateway member first decrypts the message with outer group key and then encrypts with subgroup key multicasts it to all members in the subgroup. Each member in the subgroup receives the encrypted message and performs the decryption using subgroup key and gets the original message. In this way the region-based group key agreement protocol performs the communication. The communication operation is as follows.

Source Member $\xrightarrow{E_{KR}[\text{Message}] \ \& \ \text{Multicast}}$ Gateway Member

Gateway Member $\xrightarrow{D_{KR}[E_{KR}[\text{Message}]]}$ Original Message

Gateway Member $\xrightarrow{E_{KG}[\text{Message}] \ \& \ \text{Multicast}}$ Gateway Member [Among Subgroup]

Gateway Member $\xrightarrow{D_{KG}[E_{KG}[\text{Message}]]}$ Original Message

Gateway Member $\xrightarrow{E_{KR}[\text{Message}] \ \& \ \text{Multicast}}$ Destination Member

Destination Member $\xrightarrow{D_{KR}[E_{KR}[\text{Message}]]}$ Original Message

E. Applying *Group Diffie-Hellman Key Agreement*
1. Member Join

User A and user B are going to exchange their keys(figure.6): Take **g = 5** and **p = 32713**. A's private key is nA = 76182, so A's public key PA =30754, B's private key is nB = 43310,so B's public key PB =5984. The group key is computed (Fig.[6].) User A sends its public key 30754 to user B, and then user B computes their Subgroup key as nB (A's Public key) = **16972**. User B sends its public key 5984 to User A, and then User A computes their Subgroup key as nA(B's Public key)= **16972**

GK - 16972
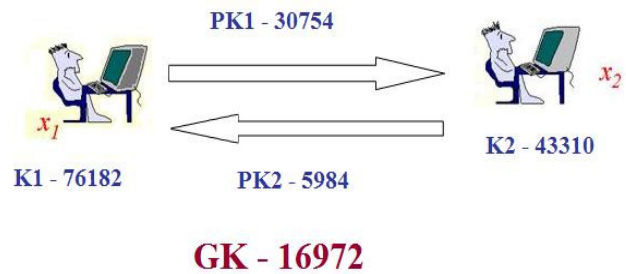
Figure.6.User-A & User –B Join the Group.

When User C is going to join in the group, C's private key becomes nC= 30561. Now, User C becomes a Subgroup Controller. Then, the key updating process will begin as follows: The previous Subgroup Controller User B sends the intermediate key as (B's Public key $ A's Public Key $ Group key of A&B)= (5984 $ 30754 $ 16972) User C separates the intermediate key as B's Public key, A's Public Key and Group key of A&B=5984 , 30754 and 16972.Then, User C generates the new Subgroup key as nC (Subgroup key of A&B)= $16972^{30561} \bmod 32713 =$ **25404**. Then, User C broadcasts the intermediate key to User A and User B. That intermediate key is ((Public key of B & C) $ (Public key of A & C)) = (25090 $1369). Now, User B extracts the value of public key of A & C from the value sent by User C. Then User B compute the new Subgroup key as follows: nB (Public key of A&C)= $1369^{43310} \bmod 32713 =$ **25404 .** Similarly, User A extracts the value of public key of B & C from intermediate key, sent by User C. Then User A compute the new Subgroup key as follows: nA (public key of B&C) = $25090^{76182} \bmod 32713 =$ **25404**. Therefore, New Subgroup Key of A, B and C = **25404** is as shown in the figure.7.
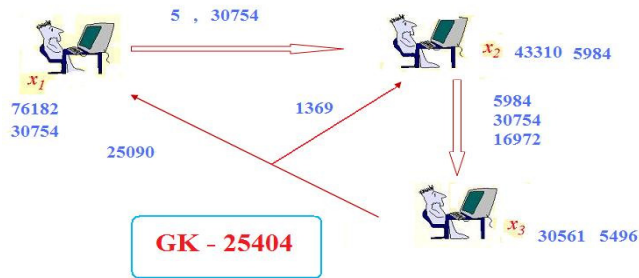
Figure .7. User- C Join in the Group.

The same procedure is followed when User D joins as shown in the Fig.8.
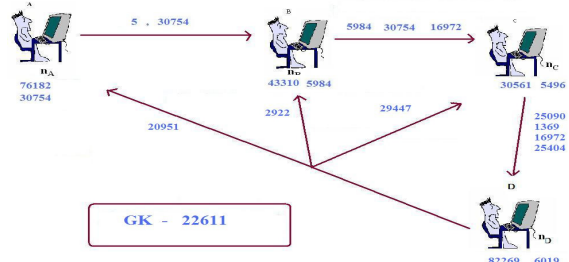
Figure.8. User-D Join in the Group.

2. Member Leave

When a user leaves (Fig.9.) from the Subgroup, then the Subgroup controller changes its private key. After that, it broadcasts its new public key value to all users in the Subgroup. Then, new Subgroup key will be generated. Let us consider, User B is going to leave, then the Subgroup Controller D changes its private key nD' =12513 ,so public key of User A & User C =11296,139)$26470. Then the new Subgroup Key generated is = 25404^{12513} mod 32713 = **5903.** Then, User A & User C computes the new Subgroup Key by using new public key. Therefore, the new Subgroup Key is **5903**.
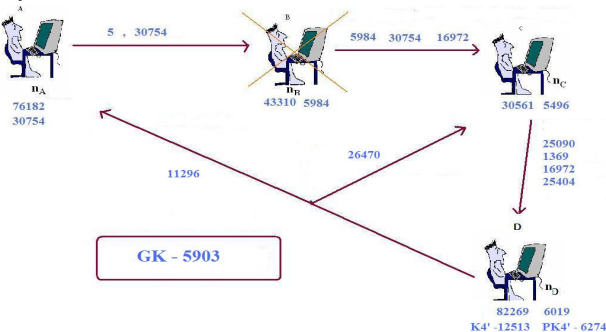
Figure.9. User –B leave from the Group.

3. Group Controller Leave

When a Subgroup controller leaves (Fig.10.) from the group, then the previous Subgroup controller changes its private key. After that, it broadcasts its new public key value to all users in the group. Then, new Subgroup key will be generated. Let us consider that the Subgroup Controller User D is going to leave, then the previous Subgroup controller User C act as Subgroup Controller and changes its private key nc' = 54170, and computes the public key of B&C $ A&C =

17618$14156. Then the new Subgroup Key generated is = 16972^{54170} mod 32713 = **27086**. Then, User A & User B compute the new Subgroup Key by using new public key. Therefore, the new Subgroup Key is **27086.**
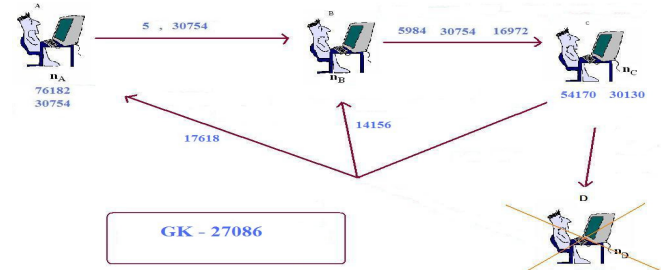
Figure.10. Group Controller Leave from the group.

F. Tree-based Group Diffie-Hellman Protocol

In the proposed protocol (Fig.11.), Tree-based group Diffie-Hellman (TGDH), a binary tree is used to organize group members. The nodes are denoted as < l, v >, where 0 <= v <= 2^l – 1 since each level l hosts at most 2^l nodes. Each node < l, v > is associated with the key K<l,v> and the blinded key BK<l,v> = F(K<l,v>) where the function f (.) is modular exponentiation in prime order groups, that is, $f(k) = α^k$ mod p (equivalent to the Diffie–Hellman protocol. Assuming a leaf node < l, v > hosts the member Mi, the node < l, v > has M_i's session random key K<l,v>. Furthermore, the member Mi at node < l. v > knows every key in the key-path from < l, v > to < 0, 0 >. Every key K<l,v> is computed recursively as follows:
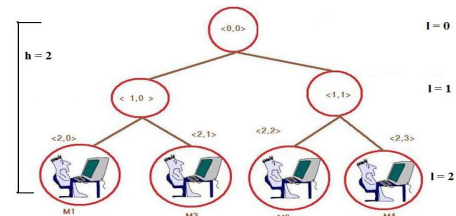
Figure.11. Key Tree.

$$K_{<l,v>} = K_{<l+1,2v>}BK_{<l+1,2v+1>} \bmod p$$
$$= K_{<l+1,2v+1>}BK_{<l+1,2v>} \bmod p$$
$$= K_{<l+1,2v>}K_{<l+1,2v+1>} \bmod p$$
$$= F(K_{<l+1,2v>}K_{<l+1,2v+1>})$$

It is not necessary for the blind key BK<l,v> of each node to be reversible. Thus, simply use the x-coordinate of K<l,v> as the blind key. The group session key can be derived from K<0,0>. Each time when there is member join/leave, the outer group controller node calculates the group session key first and then broadcasts the new blind keys to the entire group and finally the remaining group members can generate the group session key.

1. When node M_1&M_2 Join the group.

User M_1 and User M_2 are going to exchange their keys: Take **g = 5** and **p = 32713**. User M_1's private key is

79342, so M$_1$'s public key is 16678. User M$_2$'s private key is 85271, so M$_2$'s public key is **27214**. The Outer Group key is computed (Figure.12) as User M$_1$ sends its public key 16678 to user M$_2$, the User M$_2$ computes their group key as **12430**. Similarly, User M$_2$ sends its public key **27214** to user M$_1$, and then the user M$_1$ computes their group key as **12430**. Here, Outer Group controller is User M$_2$.
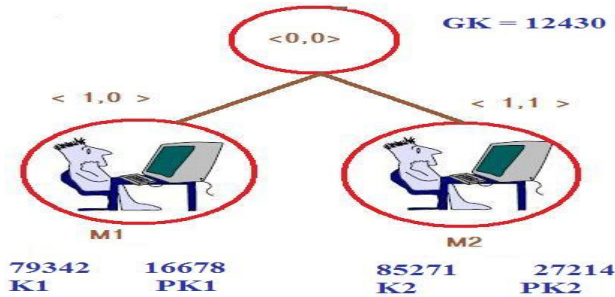
Figure.12. User M$_1$ & M$_2$ Join the Group

2. When 3rd node Join

When User M$_3$ joins the group, the old Outer group controller M$_2$ changes its private key value from 85271 to 17258 and passes the public key value and tree to User M$_3$. Now, M$_3$ becomes new Outer group controller. Then, M$_3$ generates the public key 7866) from its private key as 69816 and computes the Outer group key as 23793 shown in Figure.13. M$_3$ sends Tree and public key to all users. Now, user M$_1$ and M$_2$ compute their group key. The same procedure is followed by joining the User M$_4$ as shown in Fig.14.
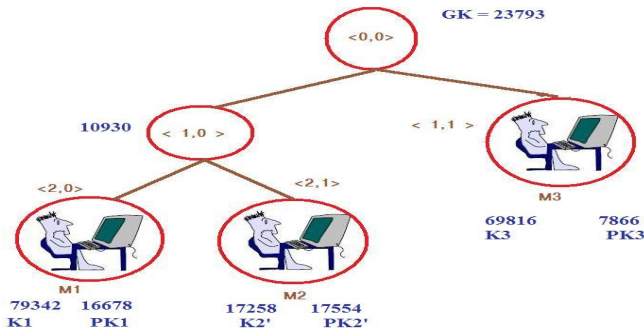
Figure.13. User M$_3$ Join the Group
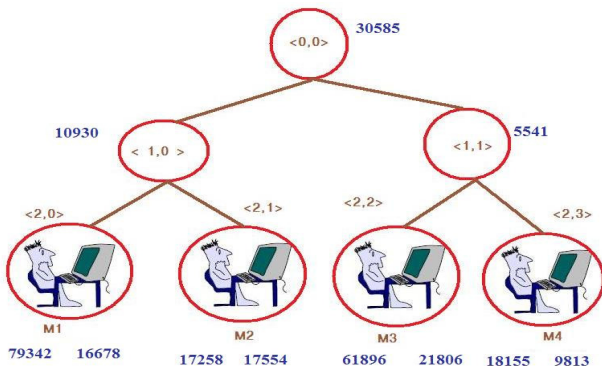
Figure.14. User M$_4$ Join the group

3. Leave Protocol

There are two types of leave, 1.Gateway Member Leave and 2.Outer Group Controller Leave

a). Gateway Member Leave

When user M$_3$ leaves (Figure.15) the Outer group, then the Outer Group controller changes its private key 18155 to55181 and outer group key is recalculated as 13151. After that, it broadcasts its Tree and public key value to all users in the Outer group. Then, the new Outer group key will be generated by the remaining users.
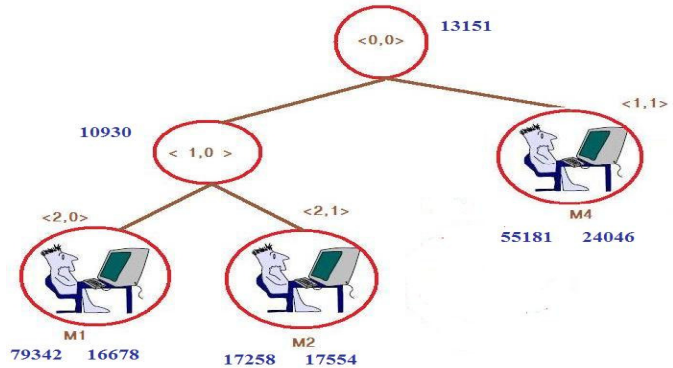
Figure.15. User M$_3$ Leave from the Group

b). When an Outer Group Controller Leaves

When an Outer Group Controller Leaves (Figure.16) from the group, then its sibling act as a New Outer Group Controller and changes its private key value 61896 to 98989 and recalculates the outer group key as 23257. After that, it broadcast its Tree and public key value to all users in the Outer group. Then, the new Outer group key will be generated by the remaining users.
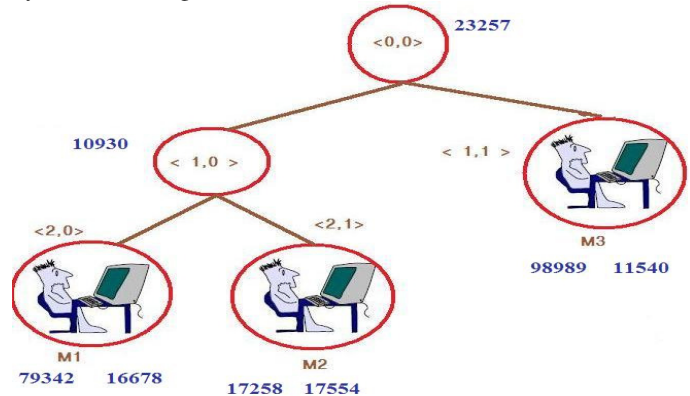
Figure.16. Outer Group Controller Leave from the Group

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were conducted on sixteen Laptops running on a 2.4 GHz Pentium CPU with 2GB of memory and 802.11 b/g 108 Mbps Super G PCI wireless cards with Atheros chipset. To test this project in a more realistic environment, the implementation is done by using Net beans IDE 6.1, in an ad-hoc network where users can securely share their data. This project integrates with a peer-to-peer (P2P) communication module that is able to communicate and share their messages with other users in the network.

The following figures are organized as follows. As described in Section III. Figure 17 shows the sub group key of user 1, 2, 3&4 in RBGKA for SGC using Group Diffie-Hellman. Figure 18 shows the sub group key after User- 2 leaves in the subgroup. Figure 19 shows the sub group key after the subgroup controller leaves in RBGKA for SGC using GDH.

Figure 20 shows the Outer group key of user M1 and M2 for RBGKA for SGC using TGDH. Similarly, figure 21 and 22 shows the outer group key of User M3 and M4 join in the outer group. Figure 23 shows the group key after the user M3 leaves in RBGKA. Figure 24 shows the outer group key after the outer group controller leaves in RBGKA.

Figure 20. Group Key of User M_1&M_2

Figure.17. Group Key of User 1, 2, 3&4
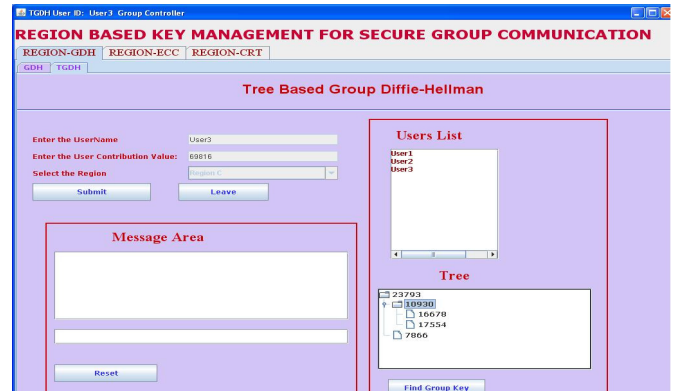
Figure 21. Group Key of User M_1, M2&M_3
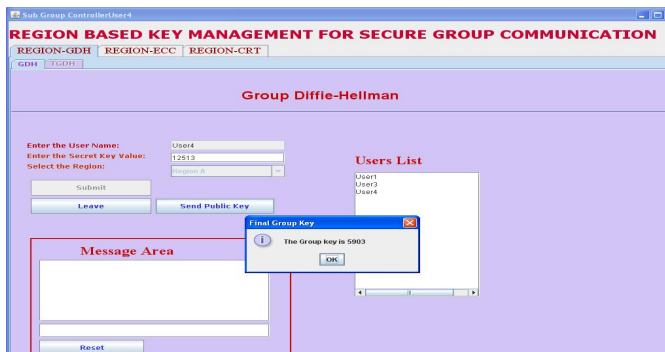
Figure.18. Group Key after User2 Leave

Figure 22. Group Key of User M_1, M2, M3 & M_4
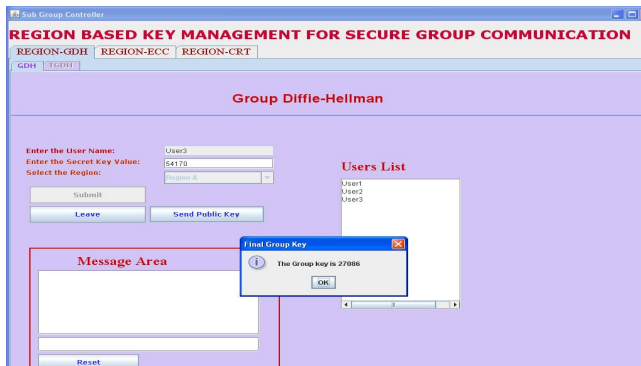
Figure.19. Group Key after Sub group controller Leave
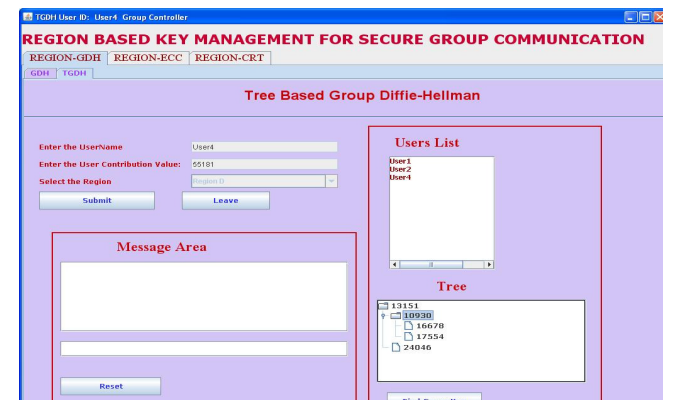
Figure 23. Group Key after M_3 Leave
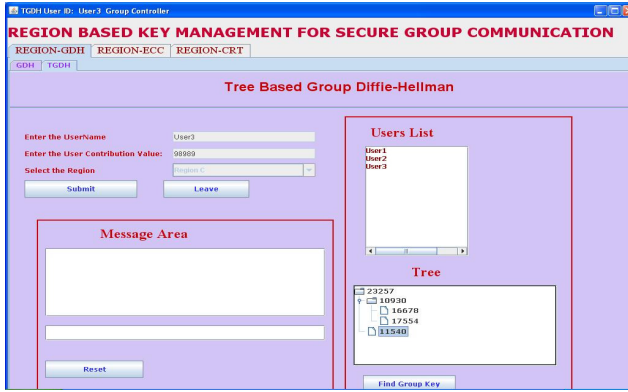
Figure.24. Group Key after Group Controller Leave

V. PERFORMANCE ANALYSIS

A. Memory Costs:

Memory cost is directly proportional to the number of members in case of TGDH and GDH. So, when the members go on increasing, TGDH and GDH occupy large memory. But in our proposed Region-Based approach, it consumes very less memory even when the members get increased. This is shown in the figure 25 and table.1.

Table 1:   Memory Cost

Protocol		Keys	Public Key Values
GDH	Concretely	2	N+1
TGDH	Per(L,V)	L+1	2N-2
	Averagely	$[\log_2 N]+1$	2N-2
RBGKA (GDH& TGDH) PROTOCOL	Member	2	X+1
	Group Controller	2+M	X+2Y -1
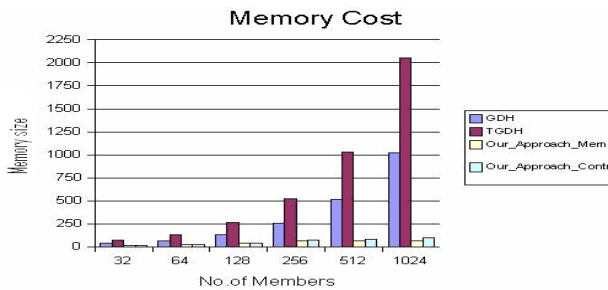
Figure 25 . Memory Cost

Consider 1024 members in a group, our approach consumes only 10% of memory comparing to GDH and 5 % of memory comparing to TGDH. Hence, we can conclude that the ratio of memory occupied is very less in our approach.

B. Communication Costs:

1. *Communication Costs – Join and Leave*

The communication cost (Table.2) depends upon the number of member joining and leaving the group. so, if there is an increase in the number of members of the group, the costs also will increase subsequently. But in our Region – Based approach, the member join/leave the subgroup is strictly restricted to a maximum of 100. In addition to that, communication of TGDH depends on trees height, balance of key tree, location of joining and leaving nodes. It also consumes more bandwidth. But our proposed approach depends only on the number of subgroup and height of tree , the communication costs get much lesser than TGDH.

Table 2: Communication and Computation Costs

Protocol Suite	Protocol	Rounds	Communication		Computation
			Unicast Size	Broadcast Size	Serial Exponentiations
GDH	Join	2	N+1	N+1	2N+1
	Leave	1	0	N-1	N-1
TGDH	Join	2	0	2N+2	3H
	Leave	1	0	2N-4	3H
Our Protocol (GDH &TGDH) — Member	Join	2	X+1	X+1	2X+1
	Leave	1	0	X-1	X-1
Our Protocol (GDH &TGDH) — Group Controller	Join	2	X+1	X+2Y+3	2X+3H+1
	Leave	1	0	X+2Y-5	X+3H-1

Where

N is the number of member in the group.
X is the number of member in the subgroup
Y is the number of Group Controller.
H is the height of the tree.
M = L+1
L is the level of the member

Considering (Figure-26) 512 members in a group, our approach consumes only 10% of Bandwidth when compare to GDH and TGDH in case of member join.
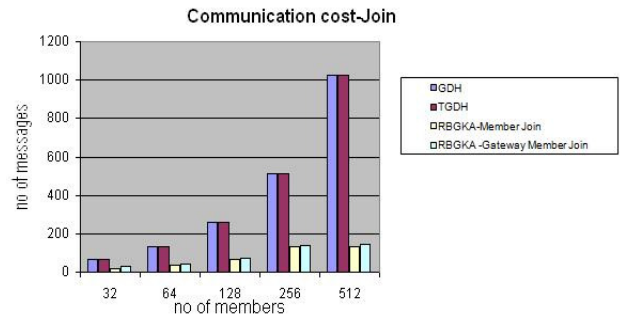
Figure 26 . Communication Cost –Join
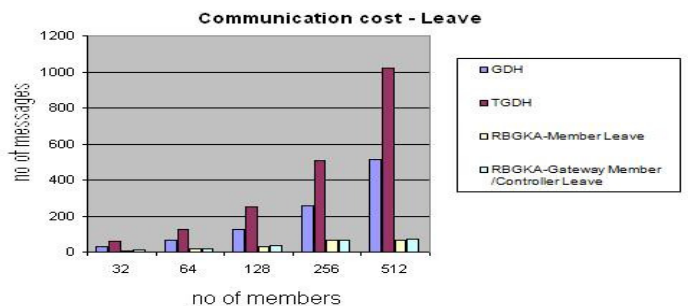
Figure 27. Communication Cost -Leave

In case of member leave, as shown in figure 27, our approach consumes 20% of Bandwidth comparing to GDH and 10% comparing to TGDH.

C. Computation Costs:

The Computational cost depends on the Serial exponentiations and the number of members joining and leaving the group. So, when the member and group size increase, the computation cost also increases significantly. Considering this fact, GDH has high computation costs as it depends on the number of members and group size. But our approach spends a little on this computation.

1.Computation Costs – Join and Leave

During member join, our approach consumes nearly 15% of serial exponentiations comparing to GDH when there are 512 members in a group. This is shown in figure 28.

Considering 512 members in a group and during member leave, our approach consumes nearly 15% of serial exponentiations when compared to GDH. Performance wise our approach leads the other two methods, even for the very large groups.
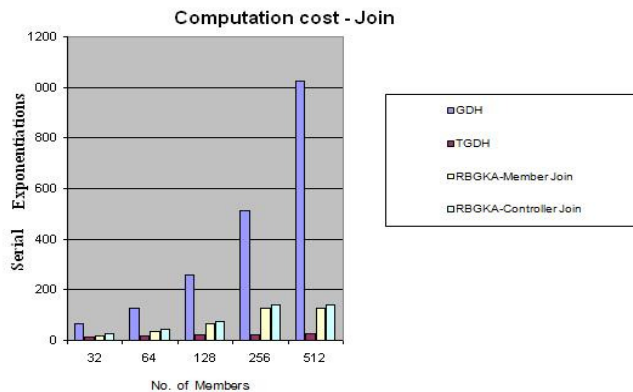
Figure 28. Computation Cost -Join
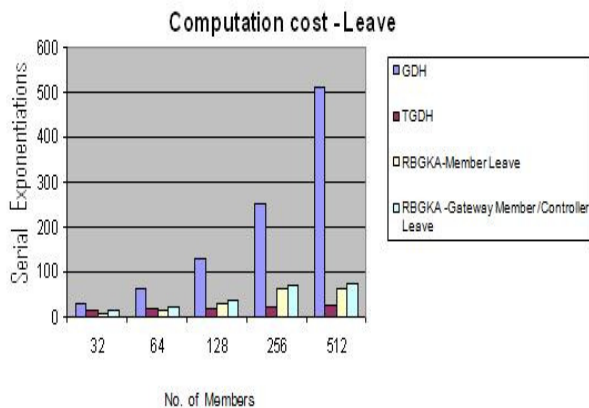
Figure 29. Computation Cost - Leave

VI.   CONCLUSION

In this paper, a region-based key agreement scheme has been proposed and implemented, which can enhance the secure group communication performance by using multiple group keys. In contrast to other existing schemes using only single key, the new proposed scheme exploits asymmetric key, i.e an Outer group Key and multiple Subgroup keys, which are generated from the proposed Region-Based key agreement algorithm. By using a set comprising an outer group key and subgroup keys a region-based scheme can be efficiently distributed for multiple secure groups. Therefore, the number of rekeying messages, computation and memory can be dramatically reduced. Compared with other schemes, the new proposed Region-Based scheme can significantly reduce the storage and communication overheads in the rekeying process, with acceptable computational overhead. It is expected that the proposed scheme can be the practical solution for secure group applications, especially for Battlefield Scenario.

REFERENCES

[1] Steiner.M, Tsudik.G, and Waidner.M, " Diffie-Hellman key distribution extended to group communication",In proc of 3rd ACM conference on computer and communication security , page 31-37 , May 1996.

[2] Steiner.M, Tsudik.G, and Waidner.M, " Cliques: A new approach to group key agreement", In proc of the 18th International conference on Distributed computing systems, pages 380-387, May 1998.

[3] Steiner.M, Tsudik.G, and Waidner.M, " Key Agreement in Dynamic Peer Groups", IEEE Trans. Parallel and Distributed Systems, vol. 11, no.8, Aug.2000.

[4] Yongdae Kim , Adrian Perrig and Gene Tsudik, " Simple and Fault-Tolerant Key Agreement for Dynamic Collaborative Groups", Proc seventh ACM conf Computer and Communication security , pages 235 -244 , Nov 2000.

[5] I. Ingemarsson , D.Tang and C.Wong, " A conference key distribution system ", IEEE Transactions on Information Theory, pages 714-720, Sept 1982.

[6] M.Burmester and Y.Desmedt , " A secure and efficient conference key distribution system", Int Advances in CRYPTOLOGY –EUROCRYPT,pages 275-286, May 1994.

[7] D. Steer, L.L. Strawczynski, W. Diffie, and M. Weiner, "A Secure Audio Teleconference System", CRYPTO'88, 1988.

[8] Yongdae Kim, Adrian Perrig, and Gene Tsudik, "Treebased group key agreement", *Cryptology ePrint Archive*, Report 2002/009, 2002.

[9] Rakesh Chandra Gangwar and Anil K. Sarje, "Complexity Analysis of Group Key Agreement Protocols for Ad Hoc Networks", 9th IEEE International Conf1`erence on Information Technology (ICIT'06)

Image Processing Algorithm

JPEG to Binary Conversion

Mansi Gupta
Dept. of Computer Sc. &
Engg.,
Lingaya's University,
Faridabad, Haryana,India
manasigupta18@gmail.com,

Meha Garg
Dept. of Computer Sc. &
Engg.,
Lingaya's University,
Faridabad, Haryana,India
mehagarg.be@gmail.com

Prateek Dhawan
Dept. of Computer Sc. &
Engg.,
Lingaya's University,
Faridabad, Haryana,India
prateek.3212@gmail.com

Abstract – **The JPEG processing algorithm works best on photographs and paintings of realistic scenes with smooth variations of tone and colour but is not well suited to files that will undergo multiple edits. The direct conversion of jpeg image into binary format is very low in efficiency. In this paper, the process of conversion of jpeg image to binary image is being done in a step by step manner, without using direct inbuilt function of jpeg to binary in MATLAB. As the binary image is used for comparison purposes, the jpeg image is converted into LAB format to make the luminance scale perceptually more uniform, so that the procedure becomes more efficient.**

Keywords: **LAB, Binary image, sign language**

I.   INTRODUCTION

JPEG (named after the Joint Photographic Experts Group who created the standard) is a commonly used method of lossy compression for photographic images. [1]

Another format is the binary format which has pixels with only two possible intensity values. They are normally displayed as black and white. Numerically, the two values are often 0 for black, and either 1 or 255 for white.

Binary images are often produced by thresholding a grayscale or color image, in order to separate an object in the image from the background. The color of the object (usually white) is referred to as the foreground color. The rest (usually black) is referred to as the background color. However, depending on the image which is to be threshold, this polarity might be inverted, in which case the object is displayed with 0 and the background is with a non-zero value. [2]

As for the LAB colour space, L* stands for luminance, a* is the red-green axis, and b* is the blue-yellow axis. The asterisks were added to differentiate CIE from another L,a,b model.[3] Although CIE L*a*b* has a large color gamut and is considered as the most accurate colour model, it is often used as a reference only or as an intermediary for colour space conversion.

II.   VARIOUS   METHODS   FOR COMPUTING BINARY IMAGE

The JPEG image can be converted into Binary image by writing codes using C# or Visual Basic.

This conversion can also be implemented by conversion of RGB into grayscale first and then into binary.

It can also be done with the help of an inbuilt function in MATLAB. The function is im2bw(RGB, level). Applying this function on the image for alphabet A generates a corresponding binary image in fig.1

Fig.1 JPEG and Binary image for alphabet A

III.   PROBLEM DEFINITION

The object is hands of the sign language useer which must be in fully in black when displayed in the binary image. The image in fig. 1 is not clear and has distortions too, i.e., the hand portion is not in black completely. This would hinder gaining

higher efficiencies in further processing of the image, if required.

With most gestures one-handed, signs maybe one-handed (ASL) or two handed (BSL).

Using colours to identify users' hands may pose problems when there are uncontrolled backgrounds [5] depicted in fig.5

Fig.5 Image for alphabet A

Few signs are often very similar (or even identical) in there manual features but differ in non-manual features (Fig. 6)

Fig. 6 Images for L, M, N and V alphabets

IV. METHODOLOGY

First the JPEG image is filtered to reduce noise and enhance the visual quality of the input image. Filtering constitutes an important part of any image processing pipeline where the final image is utilized for visual inspection or for automatic analysis. [4] This preprocessing helps increase the performance of the subsequent stages.

Fig.7 Filtered Image for alphabet A

Then the filtered image in RGB colour space is converted into LAB colour space. In LAB format, the figure can be segmented into three different

colour axis L*, a* and b*. The image is then viewed in these colour spaces and the sensitivity is tested. After testing the sensitivity, a suitable threshold value of L*, a*, b* or an appropriate combination of either a* and b* or any other colour axis is taken for formulating the binary images. There are different set of values for detecting the skin and differentiating it with the background colour.
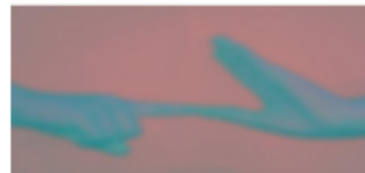
Fig.8 LAB Image for alphabet A

After the implementation of the specified values, the image can finally be converted into a binary form

Fig.9 Binary Image for alphabet A

All those pixels that have their values in this specified range are given a value of 0, i.e. white and rest all the other pixels are given a value of 1, i.e. black.

V. APPLICATIONS

The binary images can be generated for all the alphabets of BSL sign language and can be used for recognizing the alphabets. This would eliminate the need for sensors and other devices like digital gloves which have been used in sign recognition previously.
Also, it would greatly increase the efficiency for further image processing, if required, because of the near-perfect and low noise images produced.

VI. ADVANTAGES OF BINARY

- Easy to acquire: simple digital cameras can be used together with very simple frame stores, or low-cost scanners, or thresholding may be applied to grey-level images.

- Low storage: no more than 1 bit/pixel, often this can be reduced as such images are very amenable to compression (e.g. run-length coding).
- Simple processing: the algorithms are in most cases much simpler than those applied to grey-level images.

VII. DISADVANTAGES OF BINARY IMAGES

- Limited application: as the representation is only a silhouette, application is restricted to tasks where internal detail is not required as a distinguishing characteristic.
- Does not extend to 3D: the 3D nature of objects can rarely be represented by silhouettes. (The 3D equivalent of binary processing uses voxels, spatial occupancy of small cubes in 3D space).
- Specialised lighting is required for silhouettes: it is difficult to obtain reliable binary images without restricting the environment. The simplest example is an overhead projector or light box.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, the threshold determined is user-independent. It will produce exact binary images irrespective of the skin colour of the sign language user.

The JPEG images that have been taken for processing have been clicked from a fixed distance, keeping the camera position fixed too.
Future work includes removing these constraints for distance and camera position and forming clear binary images without distortions. It will focus on the extension of the developed modules in order to support larger vocabularies and enable more natural communication of the users.

REFERENCES

[1] http://en.wikipedia.org/wiki/JPEG

[2] http://www.codersource.net/csharp_color_image_to_binary.aspx

[3] http://www.answers.com/topic/cie-lab

[4] http://encyclopedia.jrank.org/articles/pages/6691/Color-Image-Filtering-and-Enhancement. html

[5] A multimodal framework for the communication of the disabled Savvas Argyropoulos 1, Konstantinos Moustakas 1, Alexey A. Karpov 2, Oya Aran 3, Dimitrios Tzovaras 1, Thanos Tsakiris 1, Giovanna Varni 4, Byungjun Kwon 5

Mansi Gupta, a final year student at Lingaya's Institute of Mgt. & Tech., Faridabad, Haryana, India. Her areas of interest include Image processing, Artificial Neural Networks, Computer organization and Operating System.

Meha Garg, a final year student at Lingaya's Institute of Mgt. & Tech., Faridabad, Haryana, India. Her areas of interest include Image processing and Artificial Neural Networks. She has published a paper in national and another in international conference during her BE level.

Prateek Dhawan, a final year student at Lingaya's Institute of Mgt. & Tech., Faridabad, Haryana, India. His areas of interest include Image processing, Artificial Neural Networks, Computer organization and Operating System.

Ontology Based Information Retrieval for

E-Tourism

G.Sudha Sadasivam
Professor,Department of CSE
PSG College of Technology
Coimbatore,India
Email id: sudhasadhasivam@yahoo.com

C.Kavitha
Senior Lecturer
PSG College of Technology
Coimbatore, India
Email id:mail2kavithak@yahoo.com

M.SaravanaPriya
PG Student
PSG College of Technology
Coimbatore, India
Email id:priyakut@gmail.com

Abstract - **This paper reports work done in the E-Tourism project. The overall goal of the project is to improve information creation, maintenance and delivery in the tourism industry by introducing semantic technologies. This paper analyzes the weakness of keyword based techniques and proposes need for semantic based intelligent information retrieval for tourism domain. The Semantic Web is an evolving development of the World Wide Web in which the meaning of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content. It also supports the transparent exchange of information and knowledge among collaborating e-business organizations. It focuses meaningful exchange of knowledge between organizations. Major challenge faced by the semantic web application is modeling of ontology and ontology based information retrieval. The software framework has been developed using Protégé tool for Travels and Tourism domain. This framework facilitates creation and maintenance of ontology. The paper also proposes two methods for information retrieval namely top down and bottom up approach. A comparison of these two approaches also presented in the paper.**

Keywords: Semantic Web, Keyword based Search Engine, Ontology, Protégé Tool, Jambalaya, Jena Agent.

I. INTRODUCTION

When surfing on the Internet, end users are increasingly in need of more powerful tools capable of searching and interpreting the vast amount of heterogeneous information available on the Web. Current Web has been designed for direct human processing, but the next-generation "Semantic Web," aims at machine-process able information[8].The Semantic Web also provides the foundation for semantic architecture to support the transparent exchange of information and knowledge among collaborating e-business organizations [2]. Recent advances in the Semantic Web technologies offer means for organizations to exchange knowledge in a meaningful way [5]. The idea allows software agents to analyze the Web on our behalf, making smart inferences that go beyond the simple linguistic analysis performed by today's search engines [5]. The applications that deliver these new online solutions are based on ontology. Ontology is basically a description of the key concepts in a given domain including the rules, properties and relationships between concepts. There are many challenges involved in implementing such an innovative new approach for online search services. Ontology modeling and ontology based information retrieval are two of the major issues faced by developers. In this paper, Ontology modeling tool Protégé and an architecture based on the tool aimed at addressing these issues are presented. The paper proposes a convenient and effective way for ontology engineer to create domain ontology enables Ontology engineer to update the ontology by adding instances and deploys effective applications and facilitates ontology based querying of Semantic Web resources.

II. PROPOSED ARCHITECTURE

Fig 1 represents a framework to support convenient and intelligent querying of Semantic Web resources for information retrieval. The key role players of this architecture include Admin, Ontology modeling tool Protégé and End user.

A. Design Steps

1. The admin or ontology engineer creates ontology by using protégé tool.

2. If any new activity is to be added to the ontology, the ontology needs to be updated. The ontology engineer updates the ontology by adding instances.

3. End user searches for web content in the same way as in a conventional search engine and issues requests using the system's GUI

4. The End users query to Jena agent and ontology will be traversed either top down or bottom up approach according to end user specification

5. The Jena agent retrieves the query result and passes the result to GUI.

6. The GUI displays the results to the end user.
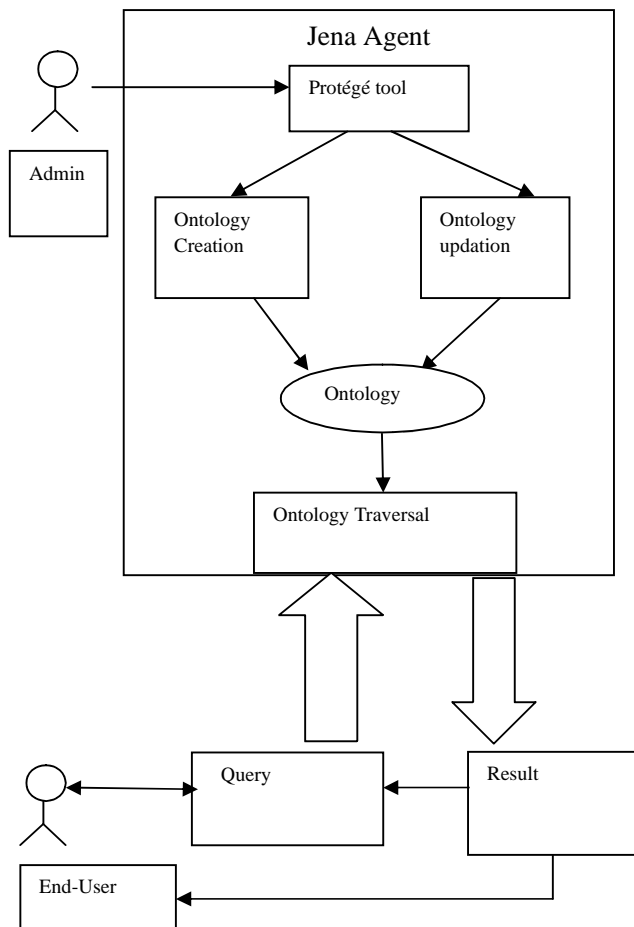
Fig. 1 System Architecture

III PROPOSED METHODOLOGY

A. *Travels and Tourism*

The Travels and Tourism Recommendation System is a travel consultancy system designed to provide budget traveling details to customer. Although travel resources on the internet are abundant, information is widely distributed among multiple travel agents. If end users want to gather information, they need to spend time searching on the internet. The results of the query are usually not accurate and sufficient. So it is necessary to design Travels and Tourism Recommendation System to help budget travelers to arrange their journey and budget [6]. In Tourism recommendation system, whenever the end user selects the source name, destination name and budget, then ontology will be traversed and meaningful result will be displayed based on top down approach. If the end user Specifies the restaurant name, accommodation name and travels name then ontology will be traversed using bottom up approach

IV IMPLEMENTATION DETAILS

Protégé is used to model ontology. It is an open source tool which is used to construct knowledge based application using ontology. Ontology is a formal explicit specification of shared conceptualization. It provides a platform for ontology engineers to create ontology and form the ontology knowledge-base. The tool displays and edits ontology in graphical mode, and can synchronously create ontology OWL [6] files as well. The work of creating ontology is realized by jambalya[9], Property Window, Individual Editor Window. According to the outline view, all the ontology objects and relative properties could be listed and displayed. The multi-layer edit view comprises two parts namely Class edit view and Property edit view. The edit view displays subclasses, instances, classes, inheritance and equivalence, mapping relation between class and instance .The edit view of Property displays properties, inheritance and equivalence relation of properties

A. *Tourism Domain Ontology Creation:*

The Protégé tool is used to create Travels and Tourism domain ontology. Fig.2 displays Travels and Tourism domain ontology created by Protégé tool using Jambalya. It has travels and tourism ontology with Travels, Restaurant, Accommodation, and Activity concepts for the cities like Mumbai, Chennai, Delhi, Hyderabad, Kolkata and Bangalore. Properties and relationship are set between each concept. Instance is created for each concept and value is assigned for each instance. Class Editor Window enables the ontology engineer to create and update the classes. Multiple siblings can be created for a same class. Based on the need, the ontology engineer can set the restrictions and comment for each classes. The ontology engineer can create a number of properties for a class using property window. Property window includes two types of properties namely Data type property, Object property. The data type property mentions the data type for each property. The ontology engineer has to specify property with corresponding subclass, range and allowable values for that property. The object poperty mentions the relationship between each class or concept. In the edit view of Property window, properties, inheritance and equivalence relation of properties are all displayed here. The ontology engineer creates number of instances or individuals for each class or concept and assign values for each instance based on data type property.
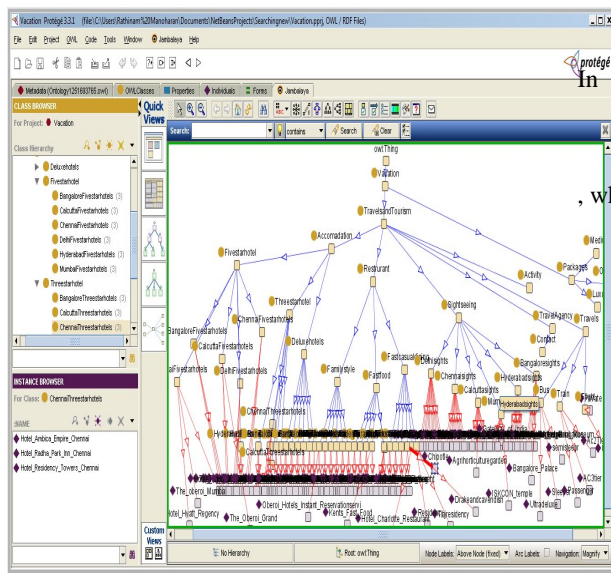
Fig. 2 Travels and Tourism Ontology

B. Tourism Domain Ontology Updation

The ontology engineer can update the ontology by adding instances.
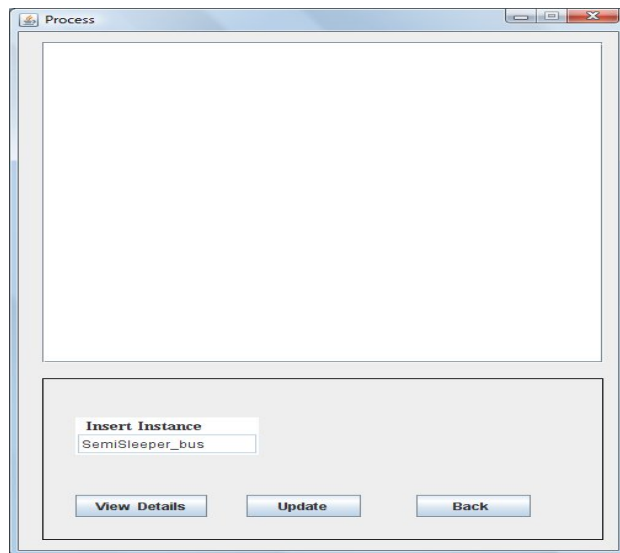
Fig. 2 Ontology updation

Fig.3 displays the ontology updation dynamically during run time. The ontology engineer has to specify the instance name with corresponding concept name to which instance is to be dynamically added. Finally ontology gets updated

C. Searching and Retrieval

In Travels and Tourism Recommendation system, when the end user issues requests the ontology will be traversed using top down approach or bottom up approach

1) Top down Approach:

In this approach the end user has to specify the source name, destination name and budget, according to budget the ontology will be traversed and results will be displayed to end user. It has 2 choices namely travel agency choice and enduser choice.

a) Travel Agency Choice: Here the end user specifies the source name, destination name and budget. The Travel agency queries the end user for his preference namely Travels or Tourism. If end user preference namely travels then details of luxury travels are extracted. If the end user preference is tourism details of tourist spots are extracted. According to end users budget, travels and tourism ontology instance weight will be added. The sum of instance weight is which is less than or equal to end user budget as results will be extracted and displayed.

Fig. 4 Travel Agency choice

Fig 4 displays the searching result based on budget estimation, preference and distance between source and destination. According to end user estimation and preference, the corresponding destination Tourist spots, Restaurant,   Accommodation and Travels details are displayed based on sum of instance weight.

b) *End User Choice:* End users are also provided with facilities to look over travels and tour spots. The end users favorite's travel, accommodation, restaurant, and tour spot and wishes to visit it in future, can be marked as his favorite spots. Fig 5 displays the information retrieval according to end user specification. The end user has to specify source name, destination name and budget category. According to budget category (luxury or medium or ordinary) the tourist spot, accommodation, restaurant and travels details are extracted.

Fig. 5 End User choice

2) Bottom up approach:

The bottom up approach is used to identify the location and category of specified activity, accommodation name, restaurant name that belong to the cities Mumbai, Chennai, Delhi, Calcutta, Bombay are displayed. Choices are available in this approach

a) Travel Agency Choice: The Fig 6 displays bottom up traversal with travel agency choice. In this approach, the travel agency queries to the end user for estimation. Based on that estimation all the instances of the sub classes are restaurant, accommodation and travels are displayed. When the end user specifies the instances this framework displays actually which destination it belongs to, type and category of restaurant, accommodation and travels.

Fig. 6 Bottom up with travel agency choice

b) *End User Choice:* The Fig 7 displays bottom up traversal with end user choice. Once the end user is provided with all the instances based on his choice the system displays destination, type and category of restaurant, accommodation and travels.

Fig. 7 Bottom up with customer choice

V PERFORMANCE EVALUATION

Finally the performance of top down approach compared with bottom up approach for information retrieval speed .The following table shown the time taken to retrieve the results.

TABLE I. PERFORMANCE COMPARISON OF TOPDOWN AND BOTTOMUPAPPROACH FOR TRAVEL AGENCY CHOICE

Performance comparison	Category	Top down approach-Enduser choice	Bottomup-Enduser choice
Time taken for Information Retrieval	Luxury	1342ms	1319ms
	Medium	1248ms	1224ms
	Low	1170ms	1143ms

GRAPH I. COMPARISON GRAPH
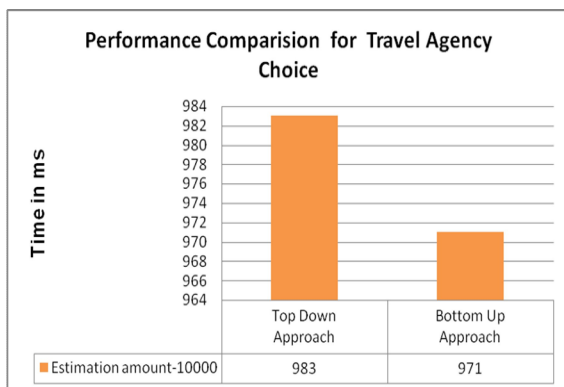
Graph I represents the performance comparison for end user choice using top down as well as bottom up approach. It shows that time taken to retrieve the information using bottom up approach is lesser than top down approach.

TABLE II. PERFORMANCE COMPARISON OF TOP DOWN AND BOTTOM UPAPPROACH FOR ENDUSER CHOICE

Top down -traveler choice	Bottom up- traveler choice
3.29 ms	3.19ms

GRAPH II. COMPARISON GRAPH
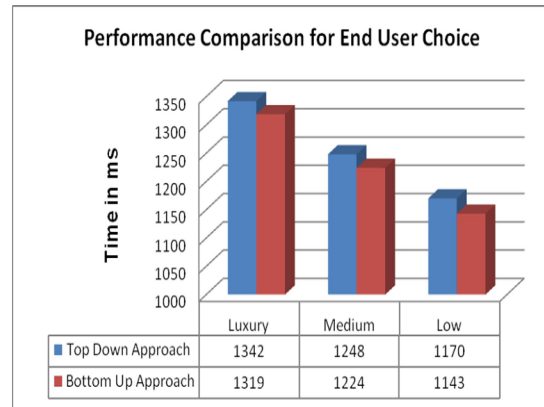
Graph II represents the performance comparison for end user choice using top down as well as bottom up approach. It shows that time taken to retrieve the information using bottom up approach is lesser than top down approach

VI CONCLUSIONS

This paper proposes the usage of ontology for travels and tourism domain. It proposes a method to create and edit ontology dynamically and a method to query for information using ontology. This paper also proposes top down and bottom up approaches to extract information from ontology. A comparison of these two approaches is also provided in this paper. When budget of travel is known and no details of instances is provided bottom up approach can not be used. Top down approach is suitable. Thus the tradeoff between top down and bottom up approaches are not only based on the performance but also on their applicability.

REFERENCES

[1] Brooke Abrahamsand Wei Dai. Architecture for Automated Annotation and Ontology Based Querying of Semantic Web Resources
[2] Konstantinos Kotis, Semantic Web Search: Perspectives and Key Technologies.Karlovassi, 83200 Samos, Greece.

[3]. Konstantinos Kotis, Dpaolo Ceravolo, Ernesto Damiani, Member, IEEE, and Marco Viviani "Bottom-Up Extraction and Trust-Based Refinement of Ontology Metadata" IEEE Transactions.

[4] Ling Li, Shengqun Tang, Lina Fang,Ruliang Xiao,Xinguo Deng,Youwei Xu,Yang Xu, Visual   Ontology Modeling Tool and Ontology Based Querying of Semantic Web Resources, 31st Annual International Computer Software and Applications Conference(COMPSAC 2007).

[5] P.H. Alesso, C. F. Smith. Developing Semantic Web Services. Canada: Wellesey MA, 2004. 165-272.

[6] P.H. Alesso, C..F. Smith, Developing Semantic Web Servces, A K Peters ltd, Wellesey MA, Canada, Date,2004, pp.165-272.

[7] Siegfried Handschuh, Steffen Staab. Authoring and annotation of web pages in CREAM. Proceedings of the 11thInternational World Wide Web Conference. USA: Honolulu, Hawaii, ACM Press, 2002. 462-473.

[8] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web - A new form of Web content that is Meaningful to computers will unleash a revolution of new possibilities," Scientific American, vol. 284, pp. 34, May 2001.

[9] The Protégé project. http://protege.stanford.edu, 2002 On Knowledge and Data Engineering, Vol. 19, No. 2, February 2007.

AUTHORS PROFILE

Dr **G Sudha Sadasivam** is working as a Professor in Department of Computer Science and Engineering in PSG College of Technology, India. Her areas of interest include Distributed Systems, Distributed Object Technology, Grid and Cloud Computing. She has published 5 books, 20 papers in referred journals and 32 papers in National and International Conferences. She has coordinated two AICTE – RPS projects in Distributed and Grid Computing areas. She is also the coordinator for PSG-Yahoo Research on Grid and Cloud computing.

Ms C Kavitha is working as a Senior Lecturer in Department of Computer Science and Engineering in PSG College of Technology, India. She is pursuing her research work in Semantics in Large scale Distriduted systems. Her areas of interest include Semantic Web Technology, Parallel Processing and Data Structures. She has published 3 papers in this area

Ms M.Saravana Priya is a PG student doing her ME – Software Engineering in CSE Department of PSG College of Technology. Her area of interest   is Semantic Web Technology. She has published 2 papers in this area

Mean – Variance parametric Model for the Classification based on Cries of Babies

Khalid Nazim S. A.,
Research Scholar.

Dr. M.B Sanjay Pande
Professor and Head,
Department of Computer Science
& Engineering, GSSSIETW,
Mysore, India.

Abstract- Cry is a feature which makes a individual to take certain care about the infant which has initiated it. It is also equally understood that cry makes a person to take certain steps. In the present work, we have tried to implement a mathematical model which can classify the cry into its cluster or group based on certain parameters based on which a cry is classified into a normal or abnormal. To corroborate the methodology we taken 17 distinguished features of cry.

The implemented mathematical model takes into account Doyle's distance to identify the required features out the 17 features for classifying the dataset. The dataset of 100 samples were taken to substantiate the efficacy of the Model.

Keywords: Cry, Doyle's distance.

I. INTRODUCTION

Is crying a normal activity is a point which opens up many queries based on method or pattern of cry. Cry is a behavior; in fact, it is a sequence of behavior patterns that is part of the larger behavioral repertoire of the infant. For the neonate and young infant, crying is the primary mode of expressing and communicating basic needs and events. It can be even defined as a signal which can be used to evaluate the neuro-respiratory and phonatory functions of the infants, which leads to the reason that cry pattern is having importance in assessing the high risk babies

Lieberman stated that it is important to study infant cry as the biological substrate of human speech involves an interplay between biological mechanisms that have other vegetative functions and neural ad anatomical mechanisms that appear to have evolved primarily for their role in facilitating human vocal communication [12].

Cry has been reported to be used as a diagnostic tool for the diagnosis of sick babies as other techniques may be invasive and may have varying amounts of risk and also require waiting until the infant is of appropriate age. With a recording and analysis of birth cry, the moment of birth itself offers data for an evaluation of the infant. Early evaluation leads to possible early detection of non normal or high risk infants which has enormous implications in the diagnosis and remediation

The model by Golub assumes that muscle control is accomplished within three levels of central nervous system processing, i.e., upper, middle and lower processors. Each of the three muscle groups important for cry production is controlled independently. Consequently the parameters that each are responsible for are likely to vary independently. Secondly, if one can pinpoint differences in the cry as caused by sub glottal (respiratory), glottal (laryngeal) or supraglottal malfunctions, then one will by able to correlate the acoustic abnormality with specific physiological and anatomical abnormalities [13].

Crying is the first tool of communication for an infant. These cries seem to be uniform, but there are a lot of differences between two infants' cries. A mother can distinguish her baby from others according to the crying. An infant cry contains a lot of information about the baby, as hunger, pain, sleepiness or boredom [4,6,7,8].Crying is a behaviour; in fact, it is a sequence of behaviour patterns that is part of the larger behavioural repertoire of the infant. For the neonate and young infant, crying is the primary mode of expressing and communicating basic needs and events. For the neonate and young infant, crying is the primary mode of expressing and communicating basic needs and events

Cry is a signal which can be used to evaluate the neuro respiratory and phonatory functions of the infants. This is the reason that cry pattern is having so much of importance in assessing the high risk babies. The abnormal infant cry is associated with chromosomal, endocrine, metabolic, and neurological disturbances, as well as malnourishment, toxicity and low birth weight i.e. infants with acoustically abnormal cries are also at long-term risk. It is possible to extract certain information from the crying sound and use it to tell whether the infant is crying due to pain, hunger or some other reason. The analysis of the infant cry involves the extraction of frequency and amplitude parameters from cry signal based on the values of these parameters infant is classified as normal or abnormal. Since cry is not one feature valued, it has many frequency and amplitude parameters. Therefore infant cry constitute the feature values in a multidimensional space [5].

For the proper assessment of the disease, a knowledge base (KB) of healthy samples with respect to that specific disease would be more useful. This then will become useful in developing a model which contrasts a test sample with the KB of healthy samples and then declares it as a healthy sample if it tallies with the KB satisfactorily; else it decides that the sample is an affected sample. This specific problem of classification can be defined as a Matching Problem. The problem will be a very focused 2-class problem, to be very precise a class and a complimentary class problem, where class refers to a healthy class and a complimentary-class refers to an unhealthy class [2].

II. METHOD

A. Subjects

A total of 59 infants were considered for the study which comprised 35 normal infants and 24 infants with high risk factor. The infants were from neonatal and sick baby wards of JSS hospital, Mysore.

Group 1: This group comprised of 35 normal infants of the age range less than 24 hrs to 1 month from the neonatal ward of JSS hospital, Mysore. They were born after 37 weeks of gestation and their birth cries were considered normal. They were born to healthy mothers who had normal delivery. The birth weight varied between 2500-3500 gms. These infants were considered to be completely healthy and normal.

Group 2: This group comprised of 24 infants of the age range less than 24 hours to 1 month form the sick baby ward of JSS hospital, Mysore and with high risk factor like prematurity, hyper bilirubinemia, jaundice, low birth weight, hypoglycemia, sibling, still birth, consanguinity, family history of speech and hearing problems and multiple risks like delayed birth cry, tachypnoea, birth asphyxia, hypertension, hypoplasia of fingers, induced labour and hypopiturarism.

B. Data collection

- Sony digital IC recorder (ICD- P320) which had an in-built microphone was used for recording the infant cries
- Laptop (Pentium dual core) with headphone and cable for line feeding of the signal was used for the analysis along with PRAAT software (version

5.0.47; Paul Boersam and David Weenink 2009; University of Amsterdam)

- Sony digital IC recorder (ICD-P320) with microphone was used to record the infant cries. It was held at a distance of approximately 5cms away from the mouth of the child. Maximum care was taken to control the noise in the room and constant intensity level was maintained for all the recordings. Thus cry samples of all the 59 infants were recorded.

C. Acoustical Processing

The acoustical analysis is the process through which the acoustical features are extracted from the crying wave; the process also implies the application of normalization and filtering techniques. By using PRAAT software the goal is to describe the signal in terms of some of its fundamental components. Input to the PRAAT is a cry signal, and its output is a vector of features that characterizes the key elements of the cry's sound wave.

We have constructed Knowledge Base employing the features of healthy samples of infant cries by removing the out layer values, which is presented in Table 1. Obviously the strength of the knowledge derived depends upon the size m. It is based on Mean (μ) and Variance (σ^2) parameters of the features of the samples. The knowledge base consists of a pair of parameters- mean (μ) and variance (σ^2), for each feature of a set of healthy samples. Generally in supervised classification, the feature values are compared with the mean values of the feature set of control samples, and subsequently the variance component is helpful for the analysis of error made by the classifier. A distance measure, called Doyle's distance measure is employed to quantify the distance that the test sample holds with the reference base. Doyle's distance model utilizes both mean and variance parameters to compute the distance [2].

TABLE I

Element	Input Feature	Mean ($\mu1$)	S D ($\sigma1$)	Doyle's Distance Components of Healthy Infant cry
F1: Median pitch	0.210310694	0.2129	0.1124	0.0466
F2: Mean pitch	0.343033	0.3284	0.1378	0.0579
F3: Minimum pitch	0.079012	0.2398	0.2692	0.1648
F4:Maximum pitch	0.70110348	0.7231	0.1827	0.0605
F5: Degree of voice breaks	0.11075216	0.2211	0.1661	0.1107
F6: Jitter (local)	0.289522389	0.3647	0.1806	0.0912
F7: Jitter (local, absolute)	0.142591959	0.2179	0.1710	0.0883
F8: Jitter (rap)	0.328521708	0.3674	0.1476	0.0663
F9: Jitter (ppq5)	0.336974673	0.3524	0.1426	0.0599
F10: Jitter (ddp)	0.328625606	0.3674	0.1476	0.0663
F11: Shimmer (local)	0.213191802	0.3142	0.1826	0.1043
F12:Shimmer (local, dB)	0.042289	0.1970	0.2127	0.1549
F13: Shimmer (apq3)	0.264902	0.3535	0.1706	0.0955
F14: Shimmer (apq5)	0.24904	0.3448	0.1520	0.0969
F15 :Shimmer (apq11)	0.189669	0.2718	0.1414	0.0849
F16:Shimmer (dda)	0.264983	0.3540	0.1451	0.0906
F17: Mean autocorrelation	0.678782	0.5931	0.1987	0.1017
Net Doyle's Distance Components of Healthy Infant cry				1.5412

Table 1: Doyle's distance values for the Healthy sample

size = m+ m = 70

TABLE II

Element	Doyle's Distance Components of Healthy Infant cry	Doyle's Distance Components of Unhealthy Infant cry
F1:Median pitch	0.0466	0.2309
F2 :Mean pitch	0.0579	0.2797
F3 : Minimum pitch	**0.1648**	**0.3215**
F4 : Maximum pitch	0.0605	0.0761
F5 : Degree of voice breaks	0.1107	0.0674
F6 : Jitter (local)	**0.0912**	**0.1521**
F7 : Jitter (local,	0.0883	0.0901
F8 : Jitter (rap)	0.0663	0.1526
F9: Jitter (ppq5)	0.0599	0.1469
F10: Jitter (ddp)	0.0663	0.1526
F11: Shimmer (local)	0.1043	0.9250
F12 Shimmer (local, dB)	**0.1549**	**0.7784**
F13: Shimmer (apq3)	0.0955	0.7340
F14: Shimmer (apq5)	**0.0969**	**0.9060**
F15 Shimmer (apq11)	**0.0849**	**1.0754**
F16 Shimmer (dda)	0.0906	0.7551
F17: Mean autocorrelation	0.1017	0.2180
Total Distance	**1.5412**	**7.0618**

Table 2: Comparison of distance between Healthy and Unhealthy infant cries

CONCLUSION

The randomly chosen sample from healthy knowledge base of infant cry was replicated m + m = 2m = 70 and Doyle's Distance value was computed which is tabulated in Table 1.

From the table 2 it can be easily understood that the parameters such as Minimum pitch, Jitter, Shimmer vary in case on unhealthy samples and also it is observed that the summation is approximately 4.5 times more in case of unhealthy samples further it is clear that the method of Doyle's distance will provide a insight in mining a data base since our results are in accordance with the previous researchers, who also had identified the same parameters for discrimination of the samples that is both healthy and unhealthy cry in infants.

Thus it results that higher affiliation index or Doyle's distance value because of an unhealthy sample indicates that the sample is refuted by the reference base. Therefore this method is a simple method to model a reference base in terms of Mean – Standard deviation as knowledge parameters is suggested. Doyle's distances are computed for affiliation analysis of a test sample. This work creates lot of scope for further improvements.

ACKNOWLEDGEMENT

REFERENCES

[1] Liisi Rautava , Asta Lempinen , Stina Ojala , Riitta Parkkola, "Acoustic quality of cry in very-low-birth-weight infants at the age of 1 1/2 years," March 2006.

[2] Sanjay Pande, PhD Thesis "An algorithmic model for exploratory analysis of trace elements in cognition and recognition of neurological disorders,", under the guidance of Dr. P Nagabhushan, Department of studies in computer Science. University of Mysore,2004.

[3] Kathleen Wermke, Ph.D., Christine Hauser, D.D.S., Gerda Komposch, D.D.S., Ph.D., and Angelika Stellzig, D.D.S., Ph.D. "Spectral Analysis of Prespeech Sounds (Spontaneous Cries) in Infants With Unilateral Cleft Lip and Palate (UCLP): A Pilot Study ,"July 1, 2001.

[4] R. G. Barr, B. Hopkins and J. A. Green. "Crying as a Sign, a Symptom, and a Signal", Mac Keith Press, London, 2000.

[5] M. Sc Thesis "Analysis of Infant Cry," under the Guidance of Dr. N.P Nataraja, 1998.

[6] Lummaa V., Vuorisalo T., Barr R. G. and Lehtonen L. "Why Cry? Adaptive Significance of Intensive Crying in Human Infants," Evolution and Human Behavior, vol. 19 (3), pp. 193 – 202, May 1998.

[7] Michelsson K., Christensson K., Rothganger H. and Winberg J., "Crying in separated and non-separated newborns: sound spectrographic analysis", Acta Pediatr, vol. 85 (4), pp. 471 – 475, April 1996.

[8] Gilbert H. R. and Robb M. P., "Vocal fundamental frequency characteristics of infant hunger cries: birth to 12 months, "Int J Pediatr Otorhinolaryngol, vol. 34, pp. 237 – 243 1996.

[9] Barbara F. Fuller , Maureen R. Keefe , Mary Curtin "Acoustic Analysis of Cries from Normal and Irritable Infants," Western Journal of Nursing Research, Vol. 16, No. 3, 243-253 (1994).

[10] QUICK Zoe L.; ROBB Michael P.; WOODWARD Lianne J. "Acoustic cry characteristics of infants exposed to methadone during pregnancy ," Acta pediatric   ISSN 0803-5253.

[11] Hartmut Rothganger, L. Wolfgang, auudge, E. Ludwig Grauel, "Jitter-index of the fundamental frequency of infant cry as a possible diagnostic tool to predict future developmental problems, "1990.

[12] Lieberman P., Harris K.S., Wolff P. & Russell L.H. (1971)," New born infant cry and non human primate vocalization", Journal of Speech and Hearing Research, 14(4) 710.

[13] Golub, H.L (1979)," A Physio acoustic model of the infant cry and its use for medical diagnosis and prognosis," In J.J Wolf and D.H Klatt (Eds.), Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America. Cambridge, MA., Acoustical Society of America.

Comparative Performance of Information Hiding in Vector Quantized Codebooks using LBG, KPE, KMCG and KFCG

Dr. H. B. Kekre
Senior Professor,
MPSTME,
NMIMS University,
Vile-parle(W),
Mumbai-56, India.
hbkekre@yahoo.com

Archana Athawale
Ph.D. Scholar, MPSTME,
NMIMS University,
Vileparle(W), Mumbai-56
Assistant Professor, TSEC,
Bandra(W), Mumbai-50,
India.

Tanuja K. Sarode
Ph.D. Scholar, MPSTME,
NMIMS University,
Vileparle(W), Mumbai-56
Assistant Professor, TSEC,
Bandra(W), Mumbai-50,
India.

Kalpana Sagvekar
Lecturer,
Fr. Conceicao Rodrigues
COE, Bandra(W),
Mumbai-50, India
kalpanasagvekar@gmail.com

Abstract - *In traditional VQ - data hiding schemes secret data is hidden inside index based cover image resulting in limited embedding capacity. To improve the embedding capacity as well as to have minimum distortion to carrier media, we have proposed one novel method of hiding secret data into the codebook. In this paper we have used four different algorithms Linde Buzo and Gray (LBG), Kekre's Proportionate Error (KPE), Kekre's Median Codebook Generation algorithm (KMCG) and Kekre's Fast Codebook Generation Algorithm (KFCG) to prepare codebooks. It is observed that KFCG gives minimum distortion.*

Keywords *- Reversible (lossless) data hiding, VQ, LBG, KPE, KMCG, KFCG.*

I. INTRODUCTION

Due to the digitalization of all kinds of data and the amazing development of network communication, information security over the Internet has become more and more important. The Internet is basically a giant open channel with security problems like modifications and interceptions occurring at any time in any place. Under such circumstances, quite some different approaches have been proposed in an attempt to make private communication secure. Researchers have developed schemes where the secret message is protected by getting transformed into the form of a stack of seemingly meaningless data, which only the authorized user can retransform back to its original form by way of some secret information. However, the appearance of a stack of seemingly meaningless data could be an irresistible attraction to an attacker with a desire to recover the original message. Another approach, called steganography, hides the secret message in some cover material with a common appearance to avoid suspicion. The data-hiding efficacy can be judged according to two criteria: (1) visual quality (2) payload capacity limit. The term "visual quality" here refers to the quality of the stego-image. That is to say, only a limited number of distortions within limited areas are allowed in the stego-image so that no obvious traces of modification

appear on the picture to catch malicious attackers' attention. Thereupon, the security of the secret information is ensured against detection. As for the payload capacity limit, it evaluates the power of a data-hiding scheme by checking how big the maximum amount of the secret information is that can be hidden in the cover media. Generally speaking, the larger the payload size is, the worse the stego-image visual quality will be. That is to say, in the world of data hiding, how to strike this balance and settle on an ideal robustness-capacity tradeoff is maybe the core problem to solve.

The existing schemes of data hiding can roughly be classified into the following three categories:

Spatial domain data hiding [2],[3],[4]: Data hiding of this type directly adjust image pixels in the spatial domain for data embedding. This technique is simple to implement, offering a relatively high hiding capacity. The quality of the stego image can be easily controlled. Therefore, data hiding of this type has become a well known method for image steganography.

Frequency domain data hiding [5],[6]: In this method images are first transformed into frequency domain, and then data is embedded by modifying the transformed coefficients.

Compressed domain data hiding [7],[8]: Data hiding is obtained by modifying the coefficients of the compressed code of a cover image. Since most images transmitted over Internet are in compressed format, embedding secret data into the compressed domain would provoke little suspicion.

Due to the restricted bandwidth of networks, we cannot keep up with the growing sizes of various multimedia files. Many popular image compression algorithms have been proposed to respond this problem, such as VQ [15], side match VQ (SMVQ) [16], JPEG [17], JPEG2000 [18], and so on. One of the most commonly studied image compression techniques is Vector Quantization (VQ) [19], which is an attractive choice

because of its simplicity and cost-effective implementation. Indeed, a variety of VQ techniques have been successfully applied in real applications such as speech and image coding [20], [22], VQ has faster encode/decode time along with simpler framework compared to JPEG/JPEG2000. Vector Quantization requires limited information during decoding and works best in applications in which the decoder has only limited information [21].

There are two approaches for hiding data into VQ compressed domain; either hides the covert data into index based cover image or in codebook. In this paper we have proposed a method of hiding data into codebook which is not been explored. In section II we present codebook design algorithms. Section III explains proposed search algorithm followed by Section IV in which results and evaluation is given. Section V gives conclusion.

II. VQ COMPRESSION TECHNIQUE

Vector Quantization (VQ) [9-14] is an efficient technique for data compression [31-34] and is very popular in a variety of research fields such as data hiding techniques [7,8], image segmentation [23-26], speech data compression [27], content based image retrieval CBIR [28, 29] and face recognition [30].

A. Codebook Generation Algorithms

a. Linde-Buzo-Gray (LBG) Algorithm [9], [10]

In this algorithm centroid is calculated as the first codevector for the training set. In Fig. 1 two vectors v1 & v2 are generated by using constant error addition to the codevector. Euclidean distances of all the training vectors are computed with vectors v1 & v2 and two clusters are formed based on nearest of v1 or v2. This procedure is repeated for every cluster. The drawback of this algorithm is that the cluster elongation is   –45° to horizontal axis in two dimensional cases. Resulting in inefficient clustering.
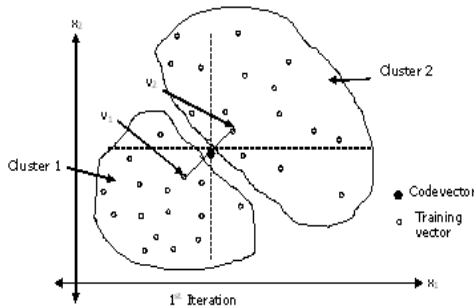
Fig.1 LBG for 2 dimensional case

b. Proportionate Error Algorithm (KPE) [11], [12]

Here proportionate error is added to the centroid to generate two vectors v1 & v2. Magnitude of elements of the centroid

decides the error ratio. Hereafter the procedure is same as that of LBG. While adding proportionate error a safe guard is also introduced so that neither v1 nor v2 go beyond the training vector space. This removes the disadvantage of the LBG. Both LBG and KPE requires 2M number of Euclidean distance computations and 2M number of comparisons where M is the total number of training vectors in every iteration to generate clusters.

c. Kekre's Median Codebook Generation Algorithm (KMCG) [13]

In this algorithm image is divided in to blocks and blocks are converted to the vectors of size k. The Fig. 2 below represents matrix T of size M x k consisting of M number of image training vectors of dimension k. Each row of the matrix is the image training vector of dimension k.

$$
T = \begin{bmatrix}
x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\
x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\
\cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot \\
x_{M,1} & x_{M,2} & \cdots & x_{M,k}
\end{bmatrix}
$$

Fig. 2. Training Vectors

The training vectors are sorted with respect to the first member of all the vectors i.e. with respect to the first column of the matrix T and the entire matrix is considered as one single cluster. The median of the matrix T is chosen (codevector) and is put into the codebook, and the size of the codebook is set to one. The matrix is then divided into two equal parts and the each of the part is then again sorted with respect to the second member of all the training vectors i.e. with respect to the second column of the matrix T and we obtain two clusters both consisting of equal number of training vectors. The median of both the parts is the picked up and written to the codebook, now the size of the codebook is increased to two consisting of two codevectors and again each part is further divided to half. Each of the above four parts obtained are sorted with respect to the third column of the matrix T and four clusters are obtained and accordingly four codevectors are obtained. The above process is repeated till we obtain the codebook of desired size. Here quick sort algorithm is used and from the results it is observed that this algorithm takes least time to generate codebook, since Euclidean distance computation is not required.

d. Kekre's Fast Codebook Generation (KFCG) Algorithm

In [14], KFCG algorithm for image data compression is proposed. This algorithm reduces the time for codebook generation. It does not use Euclidian distance for codebook generation. In this algorithm image is divided in to blocks and blocks are converted to the vectors of size k. Initially we have one cluster with the entire training vectors and the codevector C1 which is centroid.

In the first iteration of the algorithm, the clusters are formed by comparing first element of training vector with first element of code vector C1. The vector X_i is grouped into the cluster 1 if $x_{i1} < c_{11}$ otherwise vector X_i is grouped into cluster 2 as shown in Figure. 3a. where codevector dimension space is 2.

In second iteration, the cluster 1 is split into two by comparing second element x_{i2} of vector X_i belonging to cluster 1 with that of the second element of the codevector which is centroid of cluster 1. Cluster 2 is split into two by comparing the second element x_{i2} of vector X_i belonging to cluster 2 with that of the second element of the codevector which is centroid of cluster 2, as shown in Figure. 3b.

This procedure is repeated till the codebook size is reached to the size specified by user. It is observed that this algorithm gives less error as compared to LBG and requires least time to generate codebook as compared to other algorithms, as it does not require computation of Euclidian distance.
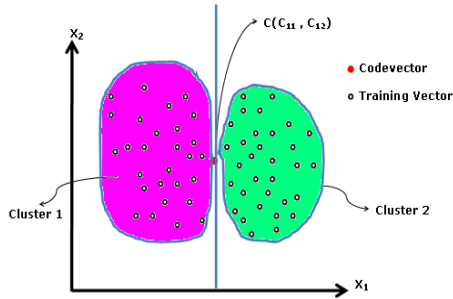
3(a). First Iteration
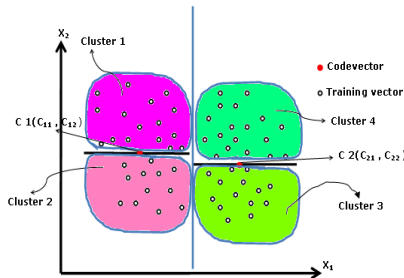
3(b) Second Iteration

Fig. 3. KFCG algorithm for 2-D case

III. PROPOSED APPROACH

In this approach, we are hiding the secret data into codebook generated using various codebook generation algorithm such as LBG[10][11], KPE[12][13], KMCG[14], KFCG[15]. There are various ways of hiding: 1bit, 2 bits, 3 bits, 4 bits & variable bits hiding.

A. Embedding Procedure
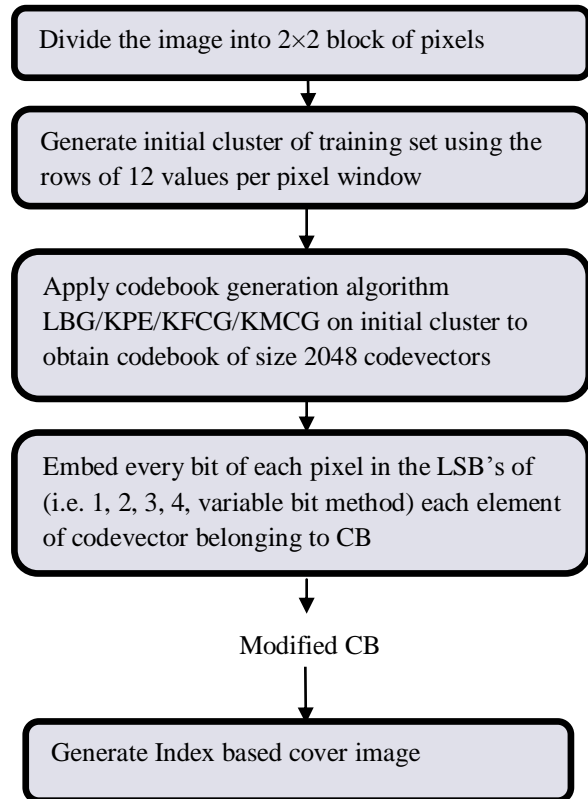
B. Extraction & Recovery Procedure
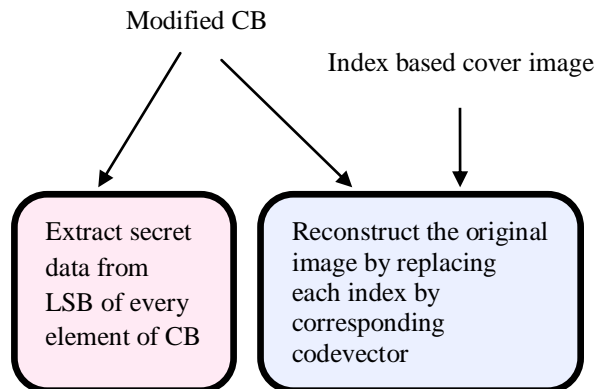
C. Variable Bit Hiding Algorithm

For variable bit hiding Kekre's algorithm [2] is used.

1. If the value of codebook vector element is in the range $240 \leq g_i \leq 255$ then we embed 4 bits of secret data into the 4 LSB's codebook vector element. This can be done by observing the 4 most significant bits (MSB's). If they are all 1's then the remaining 4 LSB's can be used for embedding data.
2. If the value of codebook vector element is in the range $224 \leq g_i \leq 239$ then we embed 3 bits of secret data. . This

can be done by observing the 3 most significant bits (MSB's). If they are all 1's then the remaining 3 LSB's can be used for embedding data.

3. If the value of codebook vector element is in the range 192≤gi≤223 then we embed 2 bits of secret data. . This can be done by observing the 2 most significant bits (MSB's). If they are all 1's then the remaining 2 LSB's can be used for embedding data.

4. If the value of codebook vector element is in the range 0≤gi≤191 we embed 1 bit of secret data.

IV. RESULTS & EVALUATIONS

In our proposed approach, we have generated codebook using LBG, KPE, KMCG and KFCG for 24 bit color image of size 256×256 shown in Fig. 4 & 5. Codebook is of size 2048×12 (i.e. 2048 code vectors each contains 12 bytes - 4 pairs of RGB). We have hidden 32×32 gray image.

Fig. 4. to Fig. 8. Shows the results of 1bit, 2bits 3bits 4bits and Variable bits using codebook obtained from LBG, KPE, KMCG and KFCG on the various cover images Bird, Fern, Puppy, Cat and Temple hiding same secrete image for fair comparison respectively.

Fig. 9. Shows the plot of Hiding Capacity versus average MSE for various hiding methods 1bit, 2bits 3bits 4bits and Variable bits on LBG, KPE, KMCG and KFCG VQ Codebooks respectively.

4 (a)
Original
Cover image
Birds.bmp

4 (b)
Secret
Message

4 (c)
Reconstructed image using
KFCG for Variable bits
Method
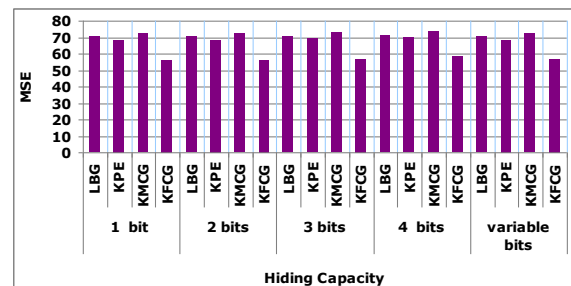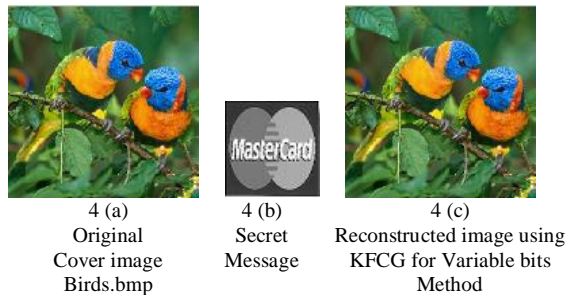
4 (d) plot of Hiding Capacity versus MSE

Fig. 4. Results of 1bit, 2bits 3bits 4bits and Variable bits on the cover image bird and secrete image shown in Fig. 4(b).

5 (a)
Original
Cover image
Fern.bmp

5 (b)
Secret
Message

5 (c)
Reconstructed image using
KFCG for Variable bits
Method
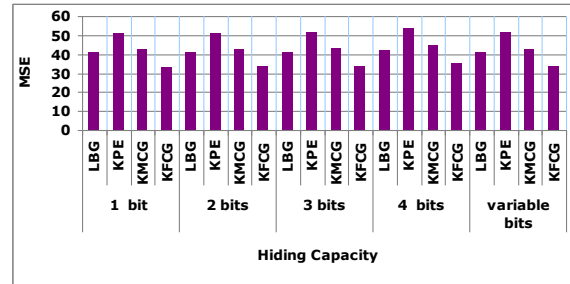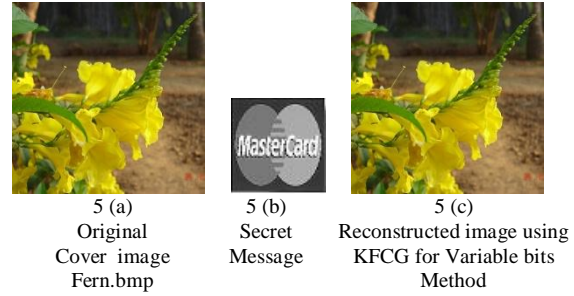
5 (d) plot of Hiding Capacity versus MSE

Fig. 5. Results of 1bit, 2bits 3bits 4bits and Variable bits on the cover image Fern shown in Fig.5(a) and secrete image shown in Fig. 5(b).

6 (a)
Original
Cover image
Puppy.bmp

6 (b)
Secret
Message

6 (c)
Reconstructed image using
KFCG for Variable bits
Method
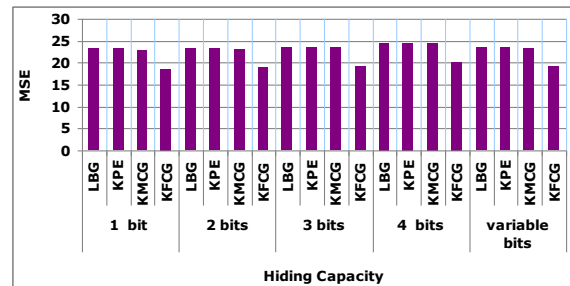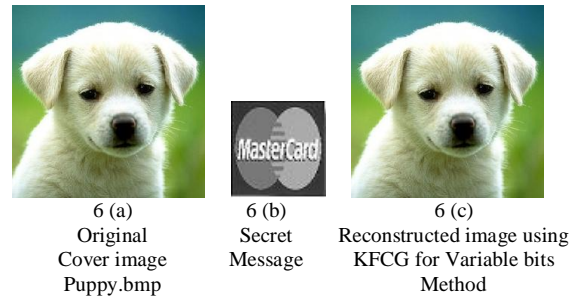
6 (d) plot of Hiding Capacity versus MSE

Fig. 6. Results of 1bit, 2bits 3bits 4bits and Variable bits on the cover image Puppy shown in Fig.6(a) and secrete image shown in Fig. 6(b).

7 (a)
Original
Cover image
Cat.bmp

7 (b)
Secret
Message

7 (c)
Reconstructed image using
KFCG for Variable bits
Method
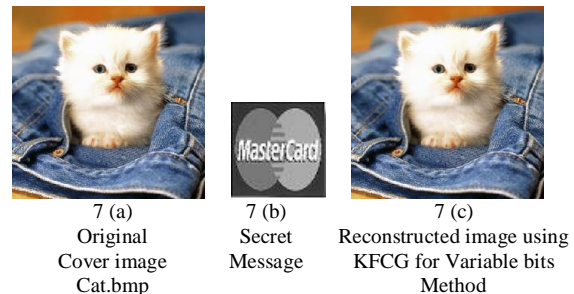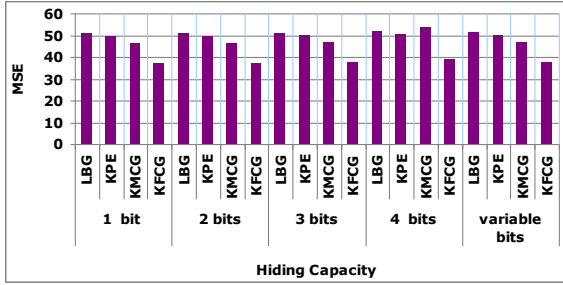
7 (d) plot of Hiding Capacity versus MSE

Fig. 7. Results of 1bit, 2bits 3bits 4bits and Variable bits on the cover image Cat shown in Fig.7(a) and secrete image shown in Fig. 7(b).

| 8 (a) Original Cover image Temple.bmp | 8 (b) Secret Message | 8 (c) Reconstructed image using KFCG for Variable bits Method |
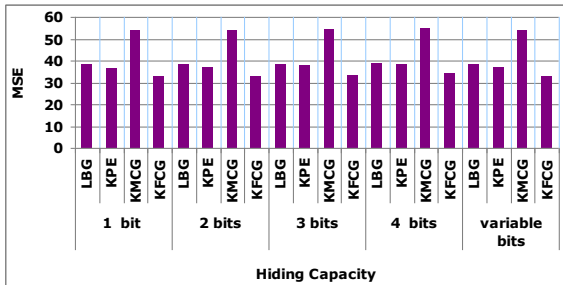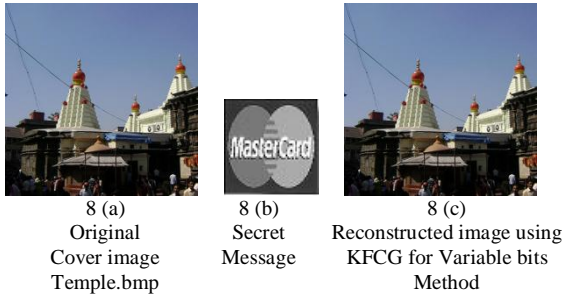
8 (d) plot of Hiding Capacity versus MSE

Fig. 8. Results of 1bit, 2bits 3bits 4bits and Variable bits on the cover image Temple shown in Fig.8(a) and secrete image shown in Fig. 8(b).
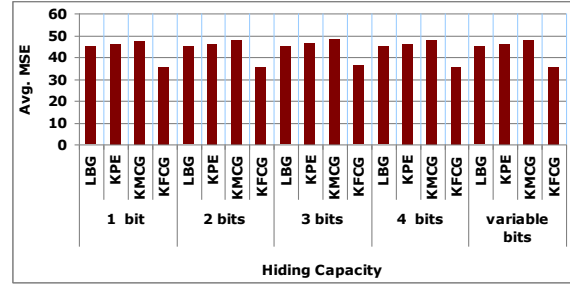
Plot of Hiding Capacity versus Avg. MSE

Fig. 9. Plot of Hiding Capacity versus average MSE for various hiding methods 1bit, 2bits 3bits 4bits and Variable bits on LBG, KPE, KMCG and KFCG VQ Codebooks respectively.

It is observed from Fig. 4 to Fig. 9. that KFCG codebook gives less MSE in all the data hiding methods 1bit, 2bits, 3bits, 4bits and varible bits as compared to LBG, KPE, and KMCG codebook. Further it is observed that varible bit method using KFCG gives the best performance.

Table 1. Shows the hiding Capacity in bits using 1 bit, 2 bits, 3 bits 4 bits, and variable bits method on LBG, KPE, KMCG and KFCG codebook of size 2048.

TABLE I. HIDING CAPACITY IN BITS USING 1 BIT, 2 BITS, 3 BITS, 4 BITS, AND VARIABLE BITS METHOD ON LBG, KPE, KMCG AND KFCG CODEBOOK OF SIZE 2048

Cover Images	Hiding Capacity in bits							
	1 bit	2 bits	3 bits	4 bits	Variable bits			
					LBG	KPE	KMCG	KFCG
Birds					28488	27202	26881	27751
Fern					27561	23891	27646	27965
Puppy	24576	49152	73728	98304	39181	38899	39962	38362
Cat					38076	37891	36364	33940
Temple					26595	26207	25545	26034

From table I it is observed that variable bits give high hiding capacity as compared to 1 bit, 2 bits, 3 bits and 4 bits embedding methods.

V. CONCLUSION

In this proposed approach the information is hidden in a vector quantized codebook by using 1,2,3,4 LSBs of the codevectors. Further a variable bit embedding is also considered which gives better embedding capacity coupled with low distortion. For preparing codebooks four different algorithms namely LBG, KPE, KMCG, KFCG are considered & their performance is considered using MSE as a parameter. It has been observed that KFCG with variable bits for hiding information gives the best performance giving mse equivalent to 2.2 bits per byte of codevectors. In addition KMCG has very low computational complexity.

REFERENCES

[1] Petitcolas, F.A.P., Anderson, R.J., and Kuhn, M.G.: 'Information hiding – a survey', Proc. IEEE, 1999, 87, (7), pp. 1062–1078 2 Swanson, M.D., Kobayashi, M., and Tewfik, A.: 'Multimedia data embedding and watermarking technologies', Proc. IEEE, 1998, 86, (6), pp. 1064–1087.

[2] H. B. Kekre, Archana Athawale and Pallavi N.Halarnkar, "Increased Capacity of Information Hiding in LSBs Method for Text and Image", International Journal of Electrical, Computer and Systems Engineering, Volume 2 Number 4. http://www.waset.org/ijecse/v2.html.

[3] H. B. Kekre, Archana Athawale and Pallavi N.Halarnkar, "Polynomial Transformation To Improve Capacity Of Cover Image For Information Hiding In Multiple LSBs", International Journal of Engineering Research and Industrial Applications (IJERIA), Ascent Publications, Volume II, March 2009, Pune.

[4] H. B. Kekre, Archana Athawale and Pallavi N.Halarnkar, "Performance Evaluation Of Pixel Value Differencing And Kekre's Modified Algorithm For Information Hiding In Images", ACM International Conference on Advances in Computing, Communication and Control (ICAC3).2009 (Uploaded on ACM Portal: http://portal.acm.org/citation.cfm?id=1523103.1523172).

[5] S.D. Lin and C.F. Chen, A Robust DCT-based Watermarking for Copyright Protection, IEEE Transactions on Consumer Electron, vol. 46, no. 3, pp. 415-421, 2000.

[6] Y.T. Wu and F.Y. Shih, Genetic algorithm based methodology for breaking the steganalytic systems, IEEE Transactions on Systems, Man and Cybernetics. Part B, vol. 36, no. 1, pp. 24-31, 2006.

[7] C. C. Chang, and C. Y. Lin, Reversible Steganography for VQ-compressed Images Using Side Matching and Relocation, IEEE Transactions on Information Forensics and Security, vol. 1, no. 4, pp. 493-501, 2006.

[8] C. C. Chang, Y. C. Chou and C. Y. Lin, Reversible Data Hiding in the VQ-Compressed Domain, IEICE Transactions on Information and Systems, vol. E90-D no. 9, pp. 1422-1429, 2007.

[9] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol. COM- 28, no. 1, pp. 84-95, 1980.

[10] A. Gersho, R.M. Gray.: 'Vector Quantization and Signal Compressio', Kluwer Academic Publishers, Boston, MA, 1991.

[11] H. B. Kekre, Tanuja K. Sarode, "New Fast Improved Codebook generation Algorithm for Color Images using Vector Quantization," International Journal of Engineering and Technology, vol.1, No.1, pp. 67-77, September 2008.

[12] H. B. Kekre, Tanuja K. Sarode, "An Efficient Fast Algorithm to Generate Codebook for Vector Quantization," First International Conference on Emerging Trends in Engineering and Technology, ICETET-2008, held at Raisoni College of Engineering, Nagpur, India, 16-18 July 2008, Avaliable at online IEEE Xplore.

[13] H. B. Kekre, Tanuja K. Sarode, "Fast Codebook Generation Algorithm for Color Images using Vector Quantization," International Journal of Computer Science and Information Technology, Vol. 1, No. 1, pp: 7-12, Jan 2009.

[14] H. B. Kekre, Tanuja K. Sarode, "New Fast Improved Codebook Generation Algorithm for Color Images using Vector Quantization", International Journal of Engg. & Tech., Vol.1, No.1, pp. 67-77, 2008.

[15] R. M. Gray, "Vector quantization," IEEE Acoust., Speech, Signal Process., vol. 1, pp. 4–29, 1984.

[16] T. Kim, "Side match and overlap match vector quantizers for images," IEEE Trans. Image Process., vol. 1, no. 4, pp. 170–185, Apr. 1992.

[17] W. B. Pennebaker and J. L. Mitchell, The JPEG Still Image Data Compression Standard. New York: Reinhold, 1993.

[18] D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals Standards and Practice. Norwell, MA: Kluwer, 2002.

[19] A. Gersho and R. M. Gray, Vector Quantization and Signal Compression. Norwell, MA: Kluwer, 1992.

[20] Z. N. Li and M. S. Drew, Fundamentals of Multimedia. Englewood Cliffs, NJ: Prentice-Hall, Oct. 2003.

[21] N. M. Nasrabadi and R. King, "Image coding using vector quantization: A review," IEEE Trans. Commun., vol. 36, no. 8, pp. 957–971, Aug. 1988.

[22] C. H. LEE, L. H. CHEN, "Fast Codeword Search Algorithm for Vector Quantization", IEE Proceedings Image Signal Processing Vol 141, No. 3 June 1994.

[23] H. B. Kekre, Tanuja K. Sarode, Bhakti Raul, "Color Image Segmentation using Kekre's Fast Codebook Generation Algorithm Based on Energy Ordering Concept", ACM International Conference on Advances in Computing, Communication and Control (ICAC3-2009), pp.: 357-362, 23-24 Jan 2009, Fr. Conceicao Rodrigous College of Engg., Mumbai. Available on ACM portal.

[24] H. B. Kekre, Tanuja K. Sarode, Bhakti Raul, "Color Image Segmentation using Kekre's Algorithm for Vector Quantization", International Journal of Computer Science (IJCS), Vol. 3, No. 4, pp.: 287-292, Fall 2008. Available: http://www.waset.org/ijcs.

[25] H. B. Kekre, Tanuja K. Sarode, Bhakti Raul, "Color Image Segmentation using Vector Quantization Techniques Based on Energy Ordering Concept" International Journal of Computing Science and Communication Technologies (IJCSCT) Volume 1, Issue 2, pp: 164-171, January 2009.

[26] H. B. Kekre, Tanuja K. Sarode, Bhakti Raul, "Color Image Segmentation Using Vector Quantization Techniques", Advances in Engineering Science Sect. C (3), pp.: 35-42, July-September 2008.

[27] H. B. Kekre, Tanuja K. Sarode, "Speech Data Compression using Vector Quantization", WASET International Journal of Computer and Information Science and Engineering (IJCISE), vol. 2, No. 4, pp.: 251-254, Fall 2008. available: http://www.waset.org/ijcise.

[28] H. B. Kekre, Ms. Tanuja K. Sarode, Sudeep D. Thepade, "Image Retrieval using Color-Texture Features from DCT on VQ Codevectors obtained by Kekre's Fast Codebook Generation", ICGST-International Journal on Graphics, Vision and Image Processing (GVIP), Volume 9, Issue 5, pp.: 1-8, September 2009. Available online at http://www.icgst.com/gvip/Volume9/Issue5/P1150921752.html.

[29] H. B. Kekre, Tanuja Sarode, Sudeep D. Thepade, "Color-Texture Feature based Image Retrieval using DCT applied on Kekre's Median Codebook", International Journal on Imaging (IJI), Volume 2, Number A09, Autumn 2009,pp. 55-65. Available online at www.ceser.res.in/iji.html (ISSN: 0974-0627).

[30] H. B. Kekre, Kamal Shah, Tanuja K. Sarode, Sudeep D. Thepade, "Performance Comparison of Vector Quantization Technique – KFCG with LBG, Existing Transforms and PCA for Face Recognition", International Journal of Information Retrieval (IJIR), Vol. 02, Issue 1, pp.: 64-71, 2009.

[31] H. B. Kekre, Tanuja K. Sarode, "2-level Vector Quantization Method for Codebook Design using Kekre's Median Codebook Generation Algorithm", Advances in Computational Sciences and Technology (ACST), ISSN 0973-6107, Volume 2 Number 2, 2009, pp. 167–178. Available online at. http://www.ripublication.com/Volume/acstv2n2.htm.

[32] H. B. Kekre, Tanuja K. Sarode, "Multilevel Vector Quantization Method for Codebook Generation", International Journal of Engineering Research and Industrial Applications (IJERIA), Volume 2, No. V, 2009, ISSN 0974-1518, pp.: 217-235. Available online at. http://www.ascent-journals.com/ijeria_contents_Vol2No5.htm.

[33] H. B. Kekre, Tanuja K. Sarode "Vector Quantized Codebook Optimization using K-Means", International Journal on Computer Science and Engineering (IJCSE) Vol.1, No. 3, 2009, pp.: 283-290, Available online at: http://journals.indexcopernicus.com/abstracted.php?level=4&id_issue=839392.

[34] H. B. Kekre, Tanuja K. Sarode, "Bi-level Vector Quantization Method for Codebook Generation", Second International Conference on Emerging Trends in Engineering and Technlogy, at

G. H. Raisoni College of Engineering, Nagpur on 16-18 December 2009, this paper will be uploaded online at IEEE Xplore.

AUTHORS PROFILE

Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm. Engineering. from Jabalpur University in 1958, M.Tech (Industrial Electronics) from IIT Bombay in 1960, M.S.Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970 He has worked as Faculty of Electrical Engg. and then HOD Computer Science and Engg. at IIT Bombay. For 13 years he was working as a professor and head in the Department of Computer Engg. at Thadomal Shahani Engineering. College, Mumbai. Now he is Senior Professor at MPSTME, SVKM's NMIMS University. He has guided 17 Ph.Ds, more than 100 M.E./M.Tech and several B.E./ B.Tech projects. His areas of interest are Digital Signal processing, Image Processing and Computer Networking. He has more than 250 papers in National / International Conferences and Journals to his credit. He was Senior Member of IEEE. Presently He is Fellow of IETE and Life Member of ISTE Recently six students working under his guidance have received best paper awards. Currently 10 research scholars are pursuing Ph.D. program under his guidance.

Ms. Archana A. Athawale has Received B.E.(Computer Engineering) degree from Walchand College of Engineering, Sangli, Shivaji University in 1996, M.E.(Computer Engineering) degree from V.J.T.I., Mumbai University in 1999, currently pursuing Ph.D. from NMIMS University, Mumbai. She has more than 10 years of experience in teaching. Presently working as - an Assistant Professor in Department of Computer Engineering at Thadomal Shahani Engineering College, Mumbai. She is a Life member of ISTE and also a member of International Association of Engineers (IAENG). Her area of interest is Image Processing, Signal Processing and Computer Graphics. She has about 30 papers in National /International Conferences/Journals to her credit.

Tanuja K. Sarode has Received Bsc.(Mathematics) from Mumbai University in 1996, Bsc.Tech.(Computer Technology) from Mumbai University in 1999, M.E. (Computer Engineering) from Mumbai University in 2004, currently Pursuing Ph.D. from Mukesh Patel School of Technology, Management and Engineering, SVKM's NMIMS University, Vile-Parle (W), Mumbai, INDIA. She has more than 10 years of experience in teaching. Currently working as Assistant Professor in Dept. of Computer Engineering at Thadomal Shahani Engineering College, Mumbai. She is life member of IETE, member International Association of Engineers (IAENG) and International Association of Computer Science and Information Technology (IACSIT), Singapore. Her areas of interest are Image Processing, Signal Processing and Computer Graphics. She has 60 papers in National /International Conferences/journal to her credit.

Kalpana R. Sagvekar has received B.E.(Computer) degree from Mumbai University with first class in 2001. Currently Perusing M.E. in Computer Engineering from University of Mumbai. She has more than 08 years of experience in teaching. Currently working as Lecturer in Computer Engineering at Fr. Conceicao Rodrigues College of Engineering, Bandra(w), Mumbai. Her areas of interest are Image Processing, Data Structure, Analysis of Algorithms, and Theoretical Computer Science. She has about 2 papers in National /International Conferences/Journals to her credit.

Registration of Brain Images using Fast Walsh Hadamard Transform

D.Sasikala [1] and R.Neelaveni [2]
[1] Research Scholar, Assistant Professor,
Bannari Amman Institute of Technology, Sathyamangalam.
Tamil Nadu-638401.
Email address : anjansasikala@gmail.com

[2] Assistant Professor,
PSG College of Technology, Coimbatore,
Tamil Nadu -641004.
Email address : rn64asok@yahoo.co.in

Abstract

A lot of image registration techniques have been developed with great significance for data analysis in medicine, astrophotography, satellite imaging and few other areas. This work proposes a method for medical image registration using Fast Walsh Hadamard transform. This algorithm registers images of the same or different modalities. Each image bit is lengthened in terms of Fast Walsh Hadamard basis functions. Each basis function is a notion of determining various aspects of local structure, e.g., horizontal edge, corner, etc. These coefficients are normalized and used as numerals in a chosen number system which allows one to form a unique number for each type of local structure. The experimental results show that Fast Walsh Hadamard transform accomplished better results than the conventional Walsh transform in the time domain. Also Fast Walsh Hadamard transform is more reliable in medical image registration consuming less time.

Keywords: Walsh Transform, Fast Walsh Hadamard Transform, Local Structure, Medical Image Registration, Normalization.

I. INTRODUCTION

Digital image processing is developing the ultimate machine that could perform the visual functions of all. It is a rapidly evolving field with growing applications in many areas of science and engineering. The main criterion of registration is to fuse the sets of data with the variations if any or with their similarities into a single data. These sets of data are acquired by sampling the same scene or object at different times or from different perspectives, in different co-ordinate systems. The purpose of registration is to visualize a single data merged with all the details about these sets of data obtained at different times or perspectives or co-ordinate systems. Such data is very essential in medicine for doctors to plan for surgery. The most common and important classes of image analysis algorithm with medical applications [1,3] are image registration and image segmentation. In Image analysis technique, the same input gives out somewhat detail description of the scene whose image is being considered. Hence the image analysis algorithms perform registration as a part of it towards producing the description. In single subject analysis, the statistical analysis is done either before or after registration. But in group analyses, it is done after registration.

Generally registration is the most difficult task, as aligning images to overlap the common features and differences if any are to be emphasized for immediate visibility to the naked eye. There is no general registration [1-17] algorithm, which can work reasonably well for all images. A suitable registration algorithm for the particular problem must be chosen or developed, as they are adhoc in nature. The algorithms can be incorporated explicitly or implicitly or even in the form of

various parameters. This step determines the success or failure of image analysis. This technique may be classified based on four different aspects given as follows: (i) the feature selection (extracting features from an image) using their similarity measures and a correspondence basis, (ii) the transformation function, (iii) the optimization procedure, and (iv) the model for processing by interpolation.

Amongst the numerous algorithms developed for image registration so far, methods based on image intensity values are particularly excellent as they are simple to automate as solutions to optimization problems. Pure translations, for example, can be calculated competently, and universally, as the maxima of the cross correlation function between two images [11] [15] [17]. Additional commands such as rotations, combined with scaling, shears, give rise to nonlinear functions which must be resolved using iterative nonlinear optimization methods [11].

In the medical imaging field, image registration is regularly used to combine the complementary and synergistic information of images attained from different modalities. A problem when registering image data is that one does not have direct access to the density functions of the image intensities. They must be estimated from the image data. A variety of image registration techniques have been used for successfully registering images that are unoccluded and generally practiced with the use of Parzen windows or normalized frequency histograms [12].

The work proposed in this paper uses Fast Walsh Hadamard Transform (FWHT) [18, 19] for image registration. The coefficients obtained are normalized to determine a unique number which in turn represents the digits in a particular range. The experiments conducted on clinical images show that proposed algorithm performed well than the conventional Walsh Transform (WT) method in medical image registration. In addition, this paper provides a comparative analysis of FWHT and WT in Medical image registration.

The remainder of the paper is ordered as follows. Section 2 provides an overview on the related work for image registration. Section 3 explains WT in image registration. Section 4 describes the proposed approach for image registration using FWHT. Section 5 illustrates the experimental results to prove the efficiency of the proposed approach in image registration and Section 6 concludes the paper with a discussion.

II. Related Work

Many discussions have been carried out previously on Image Registration. This section of paper provides a quick look on the relevant research work in image registration.

An automatic scheme using global optimization technique for retinal image registration was put forth by Matsopoulos et al. in [1]. A robust approach that estimates the affine transformation parameters necessary to register any two digital images misaligned due to rotation, scale, shear, and translation was proposed by Wolberg and Zokai in [2]. Zhu described an approach by cross-entropy optimization in [3]. Jan Kybic and Michael Unser together put forth an approach for fast elastic multidimensional intensity-based image registration with a parametric model of the deformation in [4]. Bentoutou et al. in [5] offered an automatic image registration for applications in remote sensing. A novel approach that addresses the range image registration problem for views having low overlap and which may include substantial noise for image registration was described by Silva et al. in [6]. Matungka et al. proposed an approach that involved Adaptive Polar Transform (APT) for Image registration in [7, 10]. A feature-based, fully non supervised methodology dedicated to the fast registration of medical images was described by Khaissidi et al. in [8]. Wei Pan et al. in [9] proposed a technique for image registration using Fractional Fourier Transform (FFT).

I. WALSH TRANSFORM

Orthogonal transforms expand an image into sets of orthogonal basis images each of which represents a type of local structure. Examples are the Walsh, Haar [13], etc. The coefficients of such an extension point toward the effectiveness of the occurrence of the similar structure at the particular position. If these coefficients are normalized by the dc coefficient of the expansion, i.e., the local average gray value of the image, then they measure purely the local structure independent of modality. Walsh basis functions correspond to local structure, in the form of positive or negative going horizontal or vertical edge, corner of a certain type, etc. Registration schemes based on wavelet coefficient matching do not present a general mechanism of combining the matching results across different scales.

Two images I_1 and I_2, I_1 is assumed as reference image whereas I_2 represent an image that has to be deformed to match I_1. First, consider around each pixel, excluding border pixels, a 3X3 neighborhood and compute from it, the nine Walsh coefficients (3X3 WT of a 3X3 image patch). If 'f' is the input image, the matrix of coefficients 'g' computed for it using equation (1),

$$g = (W^{-1})^T . f W^{-1} \qquad (1)$$

Matrix contains the coefficients of the expansion of the image, in terms of the basis images as in Figure.1 (a) formed by taking the vector outer products of the rows of matrix W[13]. Coefficients are denoted by $a_{00}, a_{01}, a_{02}, a_{10}, a_{11}, a_{12}, a_{20}, a_{21}, a_{22}$, in a matrix form as in Figure. 1(b) and take the value in the range [0, 9]. α_{ij} is normalization given in equation (2) makes the method robust to global levels of change of illumination. a_{00} coefficient is the local average gray value of the image, a_{ij} constructs coefficients that describes the local structure.

$$\alpha_{ij} = a_{ij} / a_{00} \qquad (2)$$

However, the information having dense features and rigid body transformation allows for plenty of redundancy in the system and makes it robust to noise and bad matches of individual pixels which effectively represent lack of local information. Construct a unique number out of eight numbers using these numbers as the digits of the unique number. The number of levels depends on the number system adopted. For decimal system, the normalized coefficients are quantized that taking integer values in the range [0, 9].
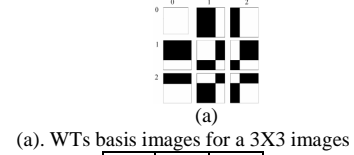
(a)
(a). WTs basis images for a 3X3 images

a_{00}	a_{01}	a_{02}
a_{10}	a_{11}	a_{12}
a_{20}	a_{21}	a_{22}

(b). Nine coefficients in matrix form
Figure 1. Walsh Transformation

In Figure 1(a) the coefficients along the first row and the first column are of equal importance, as they measure the presence of a vertical or a horizontal edge, respectively. The remaining four coefficients measure the presence of a corner. The following ordering of coefficients are used in images,

Ordering IA $\alpha_{01}, \alpha_{10}, \alpha_{20}, \alpha_{02}, \alpha_{11}, \alpha_{21}, \alpha_{12}, \alpha_{22}$

Ordering IB $\alpha_{10}, \alpha_{01}, \alpha_{02}, \alpha_{20}, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}$

Ordering IIA $\alpha_{22}, \alpha_{21}, \alpha_{12}, \alpha_{11}, \alpha_{02}, \alpha_{20}, \alpha_{10}, \alpha_{01}$

Ordering IIB $\alpha_{22}, \alpha_{12}, \alpha_{21}, \alpha_{11}, \alpha_{20}, \alpha_{02}, \alpha_{01}, \alpha_{10}$

II. PROPOSED APPROACH

A. Fast Walsh Hadamard Transform

A fast transform algorithm is seen as a sparse factorization of the transform matrix, and refers to each factor as a stage. The proposed algorithms have a regular interconnection pattern between stages, and consequently, the inputs and outputs for each stage are addressed from or to the same positions, and the factors of the decomposition, the stages, have the property of being equal between them. The 2X2 Hadamard matrix is defined as H_2 is given in equation (3)

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad (3)$$

A set of radix-R factorizations in terms of identical sparse matrices rapidly obtained from the WHT property that relates the matrix H with its inverse and is given in equation (4),

$$H_{R^n} = R^n (H_{R^n})^{-1} \qquad (4)$$

Where H_{R^n} = radix-R Walsh Hadamard transform;

R^n = radix-R factorizations;

n = input element;

The FWHT is utilized to obtain the local structure of the images. This basis function can be effectively used to obtain the digital numbers in the sense of coefficients [18] [19]. If these coefficients are normalized by the dc coefficient of the expansion, i.e., the local average gray value of the image, then they measure purely the local structure independent of modality. These numbers are then normalized to obtain the unique number that is used as feature for image registration. The implementation of FWHT readily reduces the time consumption for medical image registration when comparing the same with conventional WT technique for image registration.

III. EXPERIMENTAL RESULT

A series of experiments is performed using medical images. The tests are performed using different images of different sizes. A set of CT and magnetic resonance (MR) medical images which depict the head of the same patient is considered. The original size of these images is given as pixels. In order to remove the background parts and the head outline, the original images are cropped, creating sub-images of different dimension pixels.

In probability theory and information theory, (sometimes known as transinformation) Mutual Information between two discrete random variables is defined as the amount of information shared between the two random variables. It is a dimensionless quantity with units of bits and can be the reduction in uncertainty. High MI indicates a large reduction in uncertainty; low MI indicates a small reduction; and zero MI between two random variables means the variables are independent.

Mutual Information

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right), \quad (5)$$

- X and Y - Two discrete random variables.
- $p(x,y)$ - Joint probability distribution function of X and Y.
- $p1(x)$ and $p2(y)$ - Marginal probability distribution functions of X and Y respectively.

The Correlation Coefficient is from statistics, is a measure of how well the predicted values from a forecast model "fit" with the real-life-data. If there is no relationship between the predicted values and actual values the CC is very low. As the strength of the relationship between the predicted values and actual values increases, so does the CC. Thus higher the CC the better it is.

Correlation Coefficient

$$C(t,s;\theta) = \frac{\sum_x \sum_y [I_1^{new}(x,y) - \overline{I_1^{new}(x,y)}][I_2^{new}(x\cos\theta - y\sin\theta - t, x\sin\theta + y\cos\theta - s) - \overline{I_2^{new}(x,y)}]}{\sqrt{\sum_x \sum_y [I_1^{new}(x,y) - \overline{I_1^{new}(x,y)}]^2 \sum_x \sum_y [I_2^{new}(x\cos\theta - y\sin\theta - t, x\sin\theta + y\cos\theta - s) - \overline{I_2^{new}(x,y)}]^2}} \quad (6)$$

I_1^{new}, I_2^{new} - Two new images that differ from each other by rotation and translation only.

t, s - Shifting parameters between the two images.

θ - Rotation angle.

$\overline{I_1^{new}(x,y)}, \overline{I_2^{new}(x,y)}$ - Average structure value of the pixels in the overlapping parts of images $I_1^{new}(x,y), I_2^{new}(x,y)$ respectively.

(i) CT Sagittal Image – 432 x 427 – 41k JPEG , 36.3kB

During image registration, Figure 2,(a) the registered image of base 1 is same for both WT & FWHT. Also Figure 2.(b) shows that base 1 of both WT & FWHT gives the same difference in images.

a) Registered Image obtained for Base 1 using WT & FWHT

b) Difference in images obtained for Base 1 using WT & FWHT
Figure 2. Images obtained for Base 1 using WT & FWHT

Figure 3.(a) is the registered image of base 2 for WT. Figure 3.(b) gives the difference in images. Figure 4,(a) is the registered image of base 2 for FWHT. Figure 4.(b) shows that base 2 of FWHT gives the difference in images. Both the results are different from each other. By analyses it proves that FWHT is better when compared to the results of WT.
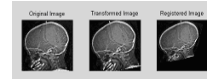
a) Registered Image obtained for Base 2 using WT

b) Difference in images obtained for Base 2 using WT
Figure 3. Images obtained for Base 2 using WT
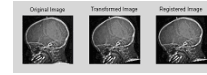
a) Registered Image obtained for Base 2 using FWHT

b) Difference in images obtained for Base 2 using FWHT
Figure 4. Images obtained for Base 2 using FWHT

a) Registered Image obtained for Base 5 using WT

b) Difference in images obtained for Base 5 using WT
Figure 5. Images obtained for Base 5 using WT
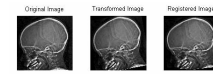
a) Registered Image obtained for Base 5 using FWHT

b) Difference in images obtained for Base 5 using FWHT
Figure 6. Images obtained for Base 5 using FWHT

a) Registered Image obtained for Base 10 using WT & FWHT

b)Difference in images obtained for Base10 using WT& FWHT
Figure 7. Images obtained for Base 10 using WT & FWHT

.(ii)   MRI T1-Registered
–   **Sagittal Image 400 x 400 –24k JPEG, 42.1kB and Frontal Image 400 x 400 – 11k JPEG, 30.9kB**

a) Registered Image obtained for Base 1 using WT & FWHT

b)Difference in images obtained for Base1 using WT& FWHT
Figure 8. Images obtained for Base 1 using WT& FWHT

MRI T1-Registered

–   **Frontal Image 400 x 400 – 11k JPEG, 30.9KB and
Sagittal Image 400 x 400 –24k JPEG, 42.1KB.**

a) Registered Image obtained for Base 1 using WT

b)Difference in images obtained for Base1 using WT

Figure 9. Images obtained for Base 1 using WT

a) Registered Image obtained for Base 1 using FWHT

b)Difference in images obtained for Base1 using FWHT

Figure 10. Images obtained for Base 1 using FWHT

MRI T1-Registered

**– Sagittal Image 400 x 400 –24k JPEG, 42.1kB and
Frontal Image 400 x 400 – 11k JPEG, 30.9kB**
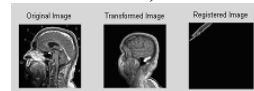
a) Registered Image obtained for Base 2 using WT

b)Difference in images obtained for Base2 using WT

Figure 11. Images obtained for Base 2 using WT

a) Registered Image obtained for Base2 using FWHT

b)Difference in images obtained for Base2 using FWHT

Figure 12. Images obtained for Base 2 using FWHT

MRI T1-Registered

**– Frontal Image 400 x 400 – 11k JPEG, 30.9kB and
Sagittal Image 400 x 400 –24k JPEG, 42.1kB.**

a) Registered Image obtained for Base 2 using WT

b)Difference in images obtained for Base2 using WT

Figure 13. Images obtained for Base 2 using WT

a) Registered Image obtained for Base2 using FWHT
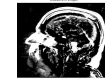
b)Difference in images obtained for Base2 using FWHT

Figure 14. Images obtained for Base 2 using FWHT

MRI T1-Registered

**– Sagittal Image 400 x 400 –24k JPEG, 42.1kB and
Frontal Image 400 x 400 – 11k JPEG, 30.9kB**

a) Registered Image obtained for Base5 using WT

b)Difference in images obtained for Base5 using WT

Figure 15. Images obtained for Base 5 using WT

a) Registered Image obtained for Base5 using FWHT
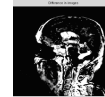
b)Difference in images obtained for Base5 using FWHT

Figure 16. Images obtained for Base 5 using FWHT

MRI T1-Registered

**– Frontal Image 400 x 400 – 11k JPEG, 30.9kB and
Sagittal Image 400 x 400 –24k JPEG, 42.1kB.**

a) Registered Image obtained for Base5 using WT

b)Difference in images obtained for Base5 using WT

Figure 17. Images obtained for Base 5 using WT

a) Registered Image obtained for Base5 using FWHT
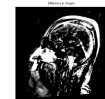
b)Difference in images obtained for Base5 using FWHT

Figure 18. Images obtained for Base 5 using FWHT

MRI T1-Registered

**– Sagittal Image 400 x 400 –24k JPEG, 42.1kB and
Frontal Image 400 x 400 – 11k JPEG, 30.9kB**

a) Registered Image obtained for Base10 using WT

b)Difference in images obtained for Base10 using WT

Figure 19. Images obtained for Base 10 using WT

a) Registered Image obtained for Base10 using FWHT
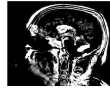
b)Difference in images obtained for Base10 using FWHT

Figure 20. Images obtained for Base10 using FWHT

MRI T1-Registered
 – Frontal Image 400 x 400 – 11k JPEG, 30.9kB and
  Sagittal Image 400 x 400 –24k JPEG, 42.1kB.
For Base 10 WT registration error occurred.

a) Registered Image obtained for Base10 using FWHT

b)Difference in images obtained for Base10 using FWHT

Figure 21. Images obtained for Base10 using FWHT

(iii) For the evaluation of the algorithm, 21 such sets of CT-MR image pairs are used.

(a) For base 1:

For MRI T2-Registered –Sagittal Image 400 x 419 - 88.8kB the results of WT and FWHT are obtained that are almost similar. Figure 22.shows the pictorial outputs from the FWHT. Even the WT produces the same output as in Figure 22.

Table 1 show the summary of all the results when a single ordering is taken into account using WT and FWHT in terms of MI. MI represents Mutual Information [16]. CC represents Correlation Coefficient. Figure 23.shows the performance comparison of WT and FWHT with respect to MI. Table 2 represents the summary of all results using conventional WT and FWHT in terms of CC. Table 3 indicates the time consumption for registering image using conventional WT and FWHT. Figure 24 represents the comparison of conventional WT and FWHT in terms of CC. Figure 25 represents the time consumption for registering image using conventional WT and FWHT.

1.

2.

3.

4.

5.

6.
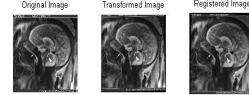
7.

8.
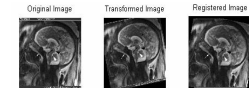
9.
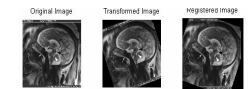
10.

11.

12.

13.
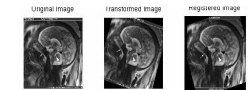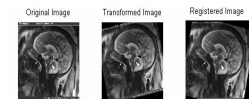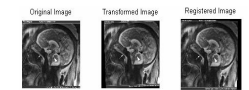
14.

15.

16.

17.
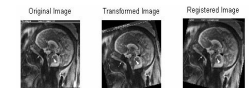
18.
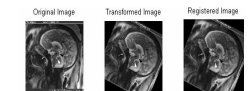
19.

20.

21.
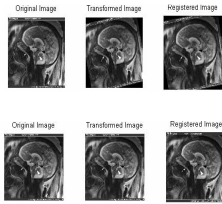
Figure 22: MRI T2-Registered –Sagittal Image 400 x 419 - 88.8kB
Using FWHT

Table 1.Represents results for WT, and FWHT using MI

S.No	X in mm	Y in mm	Angle in degrees	MI after registration for WT	MI after registration for FWHT
1	4	-10	9	0.6760	0.5759
2	-12	-7	13	0.4560	0.4580
3	5	-7	5	1.6951	0.8865
4	-14	-15	2	0.6655	0.5840
5	-8	-7	1	4.3194	2.7967
6	9	7	-7	0.8229	0.6728
7	7	-13	11	0.4955	0.4789
8	18	1	19	0.3766	0.3754
9	-17	0	-17	0.4577	0.4394
10	0	-9	12	0.5064	0.4924
11	23	-6	2	0.4982	0.4725
12	-15	5	-10	0.5726	0.5380
13	22	20	2	0.4061	0.4126
14	5	15	12	0.4790	0.4538
15	-21	16	-5	0.5023	0.5000
16	-1	19	13	0.4330	0.4239
17	5	10	-25	0.3673	0.3516
18	-3	11	25	0.3426	0.3508
19	11	-9	0	1.4506	0.9474
20	0	0	12	0.5513	0.5307
21	0	0	0	7.2952	7.2840
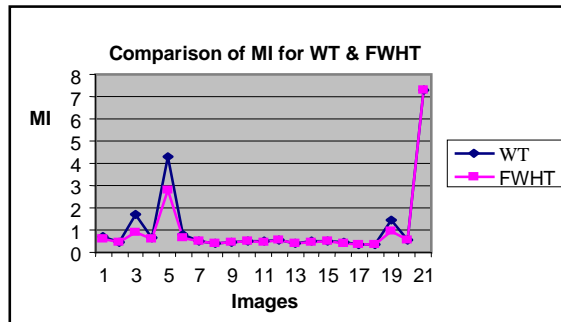
Figure 23.Comparison of WT and FWHT using MI.

Table 2.Represents results for WT, and FWHT using CC

S.No	X in mm	Y in mm	Angle in degrees	CC after registration for WT	CC after registration for FWHT
1	4	-10	9	0.4840	0.4308
2	-12	-7	13	0.3170	0.3495
3	5	-7	5	0.7801	0.6088
4	-14	-15	2	0.3876	0.4011
5	-8	-7	1	0.9425	0.9214
6	9	7	-7	0.5282	0.4889
7	7	-13	11	0.3227	0.3463
8	18	1	19	0.2199	0.2632
9	-17	0	-17	0.1774	0.1992
10	0	-9	12	0.3317	0.3658
11	23	-6	2	0.3952	0.4171
12	-15	5	-10	0.3131	0.3362
13	22	20	2	0.2667	0.3021
14	5	15	12	0.2765	0.3411
15	-21	16	-5	0.2638	0.3004
16	-1	19	13	0.2377	0.3184
17	5	10	-25	0.0987	0.1321
18	-3	11	25	0.1537	0.2109
19	11	-9	0	0.7487	0.6498
20	0	0	12	0.4324	0.4398
21	0	0	0	0.9965	0.9984

Figure 24.Comparison of WT and FWHT using CC.

Table 3.Represents Time consumption for Image Registration using WT and FWHT

S.No	X in mm	Y in mm	Angle in degrees	Elapsed Time in seconds for WT	Elapsed Time in seconds for FWHT
1	4	-10	9	108.703000	3.812000
2	-12	-7	13	105.750000	3.734000
3	5	-7	5	115.078000	3.844000
4	-14	-15	2	114.593000	3.844000
5	-8	-7	1	115.984000	3.937000
6	9	7	-7	115.078000	3.750000
7	7	-13	11	116.406000	3.766000
8	18	1	19	84.390000	3.797000
9	-17	0	-17	112.046000	3.718000
10	0	-9	12	116.562000	3.781000
11	23	-6	2	75.656000	3.719000
12	-15	5	-10	84.859000	3.813000
13	22	20	2	71.672000	3.750000
14	5	15	12	87.000000	3.781000
15	-21	16	-5	84.484000	3.797000
16	-1	19	13	89.828000	3.735000
17	5	10	-25	77.781000	3.703000
18	-3	11	25	71.766000	3.687000
19	11	-9	0	102.000000	3.907000
20	0	0	12	116.156000	3.766000
21	0	0	0	119.312000	3.766000
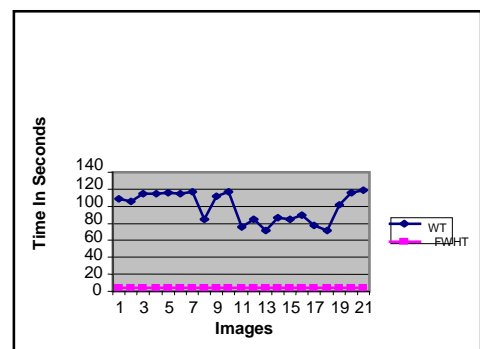
Figure 25.Comparison of WT and FWHT in terms of time
(b) For base 2:
For image 2

For image 7

For image 8

For image 9

For image 10

For image 11

For image 12
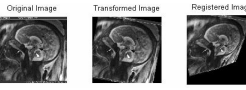
For image 13

For image 14

For image 15

For image 16

For image 17

For image 18

For image 20

Figure 26: MRI T2-Registered –Sagittal Image 400 x 419 - 88.8kB using base 2 WT

Table 4.Represents results for base 2 of WT, and FWHT using MI

S.No	X in mm	Y in mm	Angle in degrees	MI after registration for base 2 W T	MI after registration for base 2 FWHT
1	4	-10	9	4.3051	0.8614
2	-12	-7	13	0.4905	0.5424
3	5	-7	5	4.3136	4.1158
4	-14	-15	2	4.3807	4.3630

5	-8	-7	1	4.3342	4.3213
6	9	7	-7	4.3320	4.1178
7	7	-13	11	0.7509	0.6762
8	18	1	19	0.5721	0.4058
9	-17	0	-17	0.5768	0.4637
10	0	-9	12	0.6196	0.6446
11	23	-6	2	0.9336	0.7394
12	-15	5	-10	0.5874	0.7129
13	22	20	2	0.7859	0.4227
14	5	15	12	0.6063	0.5107
15	-21	16	-5	0.6566	0.6576
16	-1	19	13	0.5615	0.4770
17	5	10	-25	0.5809	0.3608
18	-3	11	25	0.5893	0.3776
19	11	-9	0	7.3026	7.3031
20	0	0	12	0.6660	0.6725
21	0	0	0	7.2931	7.2887
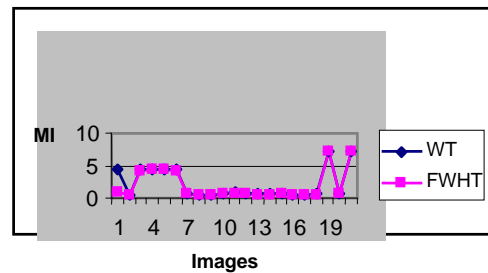
Figure 27.Comparison of base 2 of WT and FWHT using MI.

Table 5.Represents results for base 2 of WT, and FWHT using CC.

S.No	X in mm	Y in mm	Angle in degrees	CC after registration for base 2 WT	CC after registration for base 2 FWHT
1	4	-10	9	0.8347	0.5752
2	-12	-7	13	0.0276	0.4258
3	5	-7	5	0.8860	0.8864
4	-14	-15	2	0.8933	0.8934
5	-8	-7	1	0.9424	0.9426
6	9	7	-7	0.8319	0.8329
7	7	-13	11	0.1545	0.5183
8	18	1	19	0.0093	0.3054
9	-17	0	-17	-0.0163	0.2186
10	0	-9	12	0.1172	0.5127
11	23	-6	2	0.3245	0.5808
12	-15	5	-10	0.0660	0.4894
13	22	20	2	0.1492	0.3133
14	5	15	12	0.0347	0.3929
15	-21	16	-5	0.1131	0.4107
16	-1	19	13	-0.0038	0.3385
17	5	10	-25	0.0106	0.1531
18	-3	11	25	-0.0223	0.2464
19	11	-9	0	0.9006	0.8985
20	0	0	12	0.1781	0.5338
21	0	0	0	0.9908	0.9981
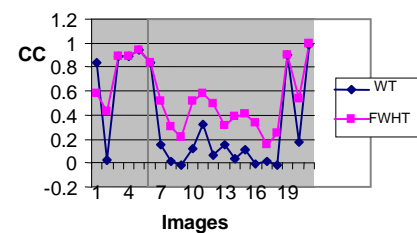
Figure 28.Comparison of base 2 of WT and FWHT using CC.

Table 6.Represents Time consumption for Image Registration using base 2 WT and FWHT

S.No	X in mm	Y in mm	Angle in degrees	Elapsed Time in seconds for base 2 WT	Elapsed Time in seconds for base 2 FWHT
1	4	-10	9	141.734000	5.485000
2	-12	-7	13	64.141000	5.125000
3	5	-7	5	144.906000	5.687000
4	-14	-15	2	143.781000	5.938000
5	-8	-7	1	144.125000	5.547000
6	9	7	-7	138.172000	5.688000
7	7	-13	11	109.234000	5.360000
8	18	1	19	67.360000	5.156000
9	-17	0	-17	54.594000	4.891000
10	0	-9	12	106.625000	5.250000
11	23	-6	2	120.781000	5.125000
12	-15	5	-10	86.922000	5.312000
13	22	20	2	89.688000	4.969000
14	5	15	12	75.234000	5.047000
15	-21	16	-5	82.547000	5.453000
16	-1	19	13	55.360000	4.984000
17	5	10	-25	55.672000	4.922000
18	-3	11	25	65.484000	4.985000
19	11	-9	0	134.703000	5.766000
20	0	0	12	110.781000	5.297000
21	0	0	0	144.875000	5.000000
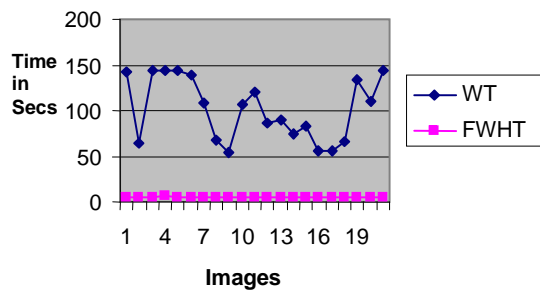
Figure 29.Comparison of base 2 of WT and FWHT in terms of time

(c) For base 5:

Table 7.Represents results for base 5 of WT, and FWHT using MI

S.No	X in mm	Y in mm	Angle in degrees	MI after registration for base 5 WT	MI after registration for base 5 FWHT
1	4	-10	9	4.3363	4.0421
2	-12	-7	13	4.3264	4.0996
3	5	-7	5	4.3136	4.1160
4	-14	-15	2	4.3802	4.3620
5	-8	-7	1	0.4197	4.3213
6	9	7	-7	4.3315	4.1368
7	7	-13	11	4.3147	4.2552
8	18	1	19	4.2660	3.6528
9	-17	0	-17	4.3429	4.3334
10	0	-9	12	4.3727	4.0351
11	23	-6	2	4.2871	4.1984
12	-15	5	-10	4.3386	4.2534
13	22	20	2	4.3571	4.0040
14	5	15	12	4.3256	4.0442
15	-21	16	-5	4.3478	3.9213
16	-1	19	13	4.3553	3.9980
17	5	10	-25	0.3670	2.7545
18	-3	11	25	error	4.2565
19	11	-9	0	7.3026	7.3028
20	0	0	12	4.3677	4.3685
21	0	0	0	7.2836	7.2923

For image 5
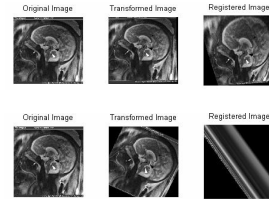
For image 17

Figure 30: MRI T2-Registered –Sagittal Image 400 x 419 - 88.8kB using base 5 WT
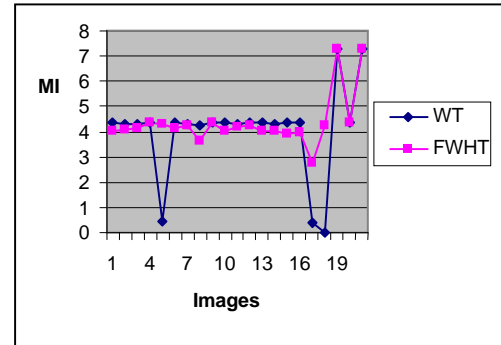
Figure 31.Comparison of base 5 of WT and FWHT using MI.

Table 8.Represents results for base 5 of WT, and FWHT using CC.

S.No	X in mm	Y in mm	Angle in degrees	CC after registration for base 5 WT	CC after registration for base 5 FWHT
1	4	-10	9	0.8342	0.8343
2	-12	-7	13	0.8459	0.8464
3	5	-7	5	0.8860	0.8864
4	-14	-15	2	0.8934	0.8934
5	-8	-7	1	0.2565	0.9426
6	9	7	-7	0.8319	0.8327
7	7	-13	11	0.7713	0.7719
8	18	1	19	0.6627	0.6648
9	-17	0	-17	0.8052	0.8054
10	0	-9	12	0.8206	0.8201
11	23	-6	2	0.8006	0.8009
12	-15	5	-10	0.8408	0.8416
13	22	20	2	0.7378	0.7381
14	5	15	12	0.7152	0.7171
15	-21	16	-5	0.8282	0.8297
16	-1	19	13	0.7201	0.7207
17	5	10	-25	0.1778	0.6876
18	-3	11	25	error	0.6574
19	11	-9	0	0.9006	0.8969
20	0	0	12	0.8226	0.8226
21	0	0	0	0.9930	0.9911
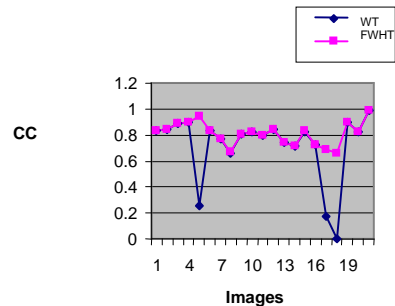
Figure 32.Comparison of base 5 of WT and FWHT using CC.

Table 9.Represents Time consumption for Image Registration using base 5 WT and FWHT

S.No	X in mm	Y in mm	Angle in degrees	Elapsed Time in seconds for base 2 WT	Elapsed Time in seconds for base 2 FWHT
1	4	-10	9	147.125000	6.828000
2	-12	-7	13	136.250000	7.297000
3	5	-7	5	141.453000	6.234000
4	-14	-15	2	136.797000	6.734000
5	-8	-7	1	138.813000	6.031000
6	9	7	-7	139.047000	6.532000
7	7	-13	11	135.640000	7.109000
8	18	1	19	131.563000	7.953000
9	-17	0	-17	132.546000	7.812000
10	0	-9	12	134.750000	7.141000
11	23	-6	2	137.797000	6.765000
12	-15	5	-10	137.234000	7.031000
13	22	20	2	134.969000	7.344000
14	5	15	12	134.609000	7.250000
15	-21	16	-5	134.985000	7.063000
16	-1	19	13	135.687000	7.406000
17	5	10	-25	76.594000	8.390000
18	-3	11	25	error	8.391000
19	11	-9	0	151.719000	6.297000
20	0	0	12	133.453000	7.109000
21	0	0	0	150.047000	5.453000
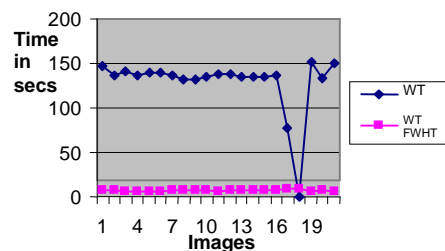
Figure 33.Comparison of base 5 of WT and FWHT in terms of time

(d) For base 10:

Table 10.Represents results for base 10 of WT, and FWHT using MI

S.No	X in mm	Y in mm	Angle in degrees	MI after registration for base 5 WT	MI after registration for base 5 FWHT
1	4	-10	9	4.3362	4.0421
2	-12	-7	13	4.3256	4.0996
3	5	-7	5	0.3304	4.1160
4	-14	-15	2	4.3722	4.3620
5	-8	-7	1	0.4205	4.3213
6	9	7	-7	4.1077	4.1368
7	7	-13	11	4.3106	4.2549
8	18	1	19	4.2627	3.6573
9	-17	0	-17	4.3432	4.3334
10	0	-9	12	4.3724	4.0351
11	23	-6	2	4.2861	4.1984
12	-15	5	-10	error	4.2534
13	22	20	2	4.3571	4.0040
14	5	15	12	4.3263	4.0442
15	-21	16	-5	4.3480	3.9213
16	-1	19	13	4.3539	3.9980
17	5	10	-25	0.3677	2.7545
18	-3	11	25	error	4.2565
19	11	-9	0	7.3026	7.3028
20	0	0	12	4.3679	4.3685
21	0	0	0	7.2892	7.2923

For Image 3

For Image 5

For Image 17
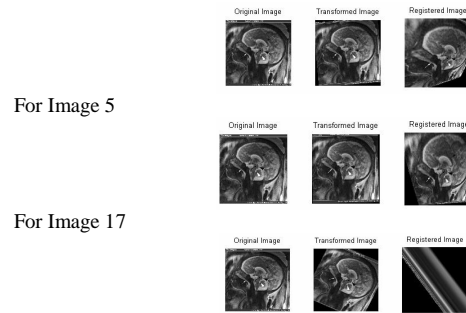
Figure 34: MRI T2-Registered –Sagittal Image 400 x 419 - 88.8kB using base 10 WT
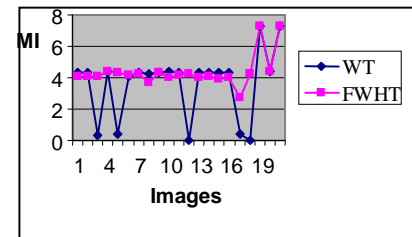
Figure 35.Comparison of base 10 of WT and FWHT using MI.

Table 11.Represents results for base 10 of WT, and FWHT using CC.

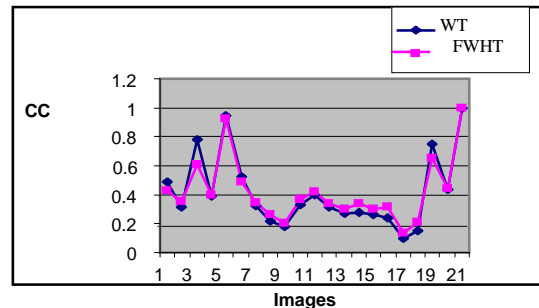| S.No	X in mm	Y in mm	Angle in degrees	CC after registration for base 5 WT	CC after registration for base 5 FWHT
1	4	-10	9	0.8342	0.8343
2	-12	-7	13	0.8459	0.8464
3	5	-7	5	0.0416	0.8864
4	-14	-15	2	0.8933	0.8934
5	-8	-7	1	0.2542	0.9426
6	9	7	-7	0.8306	0.8327
7	7	-13	11	0.7713	0.7719
8	18	1	19	0.6628	0.6645
9	-17	0	-17	0.8054	0.8054
10	0	-9	12	0.8206	0.8201
11	23	-6	2	0.8006	0.8009
12	-15	5	-10	error	0.8416
13	22	20	2	0.7378	0.7381
14	5	15	12	0.7151	0.7171
15	-21	16	-5	0.8282	0.8297
16	-1	19	13	0.7200	0.7207
17	5	10	-25	0.1778	0.6876
18	-3	11	25	error	0.6574
19	11	-9	0	0.9006	0.8969
20	0	0	12	0.8226	0.8226
21	0	0	0	0.9896	0.9911
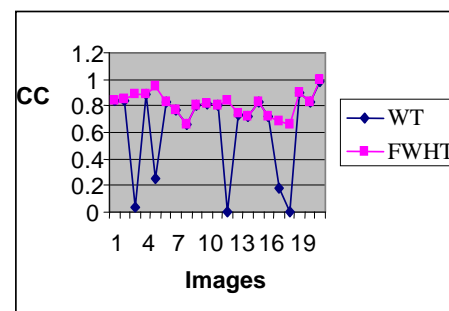
Figure 36.Comparison of base 10 of WT and FWHT using CC.

Table 12.Represents Time consumption for Image Registration using base 5 WT and FWHT

S.No	X in mm	Y in mm	Angle in degrees	Elapsed Time in seconds for base 2 WT	Elapsed Time in seconds for base 2 FWHT
1	4	-10	9	138.297000	6.829000
2	-12	-7	13	135.922000	7.328000
3	5	-7	5	133.406000	6.328000
4	-14	-15	2	136.000000	6.750000
5	-8	-7	1	141.125000	6.125000
6	9	7	-7	139.000000	6.547000
7	7	-13	11	136.703000	7.078000
8	18	1	19	131.750000	7.953000
9	-17	0	-17	132.828000	7.812000
10	0	-9	12	135.328000	7.156000
11	23	-6	2	138.907000	6.797000
12	-15	5	-10	error	6.968000
13	22	20	2	134.687000	7.344000
14	5	15	12	135.266000	7.219000
15	-21	16	-5	135.469000	7.109000
16	-1	19	13	136.782000	7.375000
17	5	10	-25	100.563000	8.343000
18	-3	11	25	error	8.390000
19	11	-9	0	142.500000	6.312000
20	0	0	12	134.797000	7.125000
21	0	0	0	149.781000	5.453000
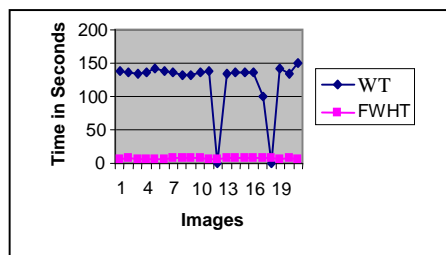
Figure 37.Comparison of base 10 of WT and FWHT in terms of time

From the above analysis it proves that the performance of the FWHT is better than the WT in terms of all the measures.

IV. CONCLUSION

This paper proposes a new algorithm for medical image registration. A Fast Walsh Hadamard Transform is proposed in this paper for medical image registration. This transform reduces the time consumption in image registration. Therefore it proves to be a better approach for medical image registration than any other conventional Walsh Transform. The coefficients obtained using this transform are then normalized to obtain the unique number. This unique number represents the local structure of an image. Moreover this unique number indicates the feature of an image for image registration. The experimental results revealed the fact the proposed algorithm using Fast Walsh Hadamard Transform performed well in image registration. The future work concentrates on further improvement in the results by using some other transforms that use correlation coefficients.

REFERENCES

[1] George K. Matsopoulos, Nicolaos A. Mouravliansky, Konstantinos K. Delibasis, and Konstantina S. Nikita, "Automatic Retinal Image Registration Scheme Using Global Optimization Techniques," IEEE Transactions on Information Technology in Biomedicine, vol. 3, no. 1, pp. 47-60, 1999.

[2] G. Wolberg, and S. Zokai, "Robust image registration using log-polar transform," Proceedings of International Conference on Image Processing, vol. 1, pp. 493-496, 2000.

[3] Yang-Ming Zhu, "Volume Image Registration by Cross-Entropy Optimization," IEEE Transactions on Medical Imaging, vol. 21, no. 2, pp. 174-180, 2002.

[4] Jan Kybic, and Michael Unser, "Fast Parametric Elastic Image Registration," IEEE Transactions on Image Processing, vol. 12, no. 11, pp. 1427-1442, 2003.

[5] Y. Bentoutou, N. Taleb, K. Kpalma, and J. Ronsin, "An Automatic Image Registration for Applications in Remote Sensing," IEEE Transactions on Geosciences and Remote Sensing, vol. 43, no. 9, pp. 2127-2137, 2005.

[6] Luciano Silva, Olga R. P. Bellon, and Kim L. Boyer, "Precision Range Image Registration Using a Robust Surface Interpenetration Measure and Enhanced Genetic Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 762-776, 2005.

[7] R. Matungka, Y. F. Zheng, and R. L. Ewing, "Image registration using Adaptive Polar Transform," 15th IEEE International Conference on Image Processing, ICIP 2008, pp. 2416-2419, 2008.

[8] G. Khaissidi, H. Tairi and A. Aarab, "A fast medical image registration using feature points," ICGST-GVIP Journal, vol. 9, no. 3, 2009.

[9] Wei Pan, Kaihuai Qin, and Yao Chen, "An Adaptable-Multilayer Fractional Fourier Transform Approach for Image Registration," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 3, pp. 400-413, 2009.

[10] R. Matungka, Y. F. Zheng, and R. L. Ewing, "Image registration using Adaptive Polar Transform," IEEE Transactions on Image Processing, vol. 18, no. 10, pp. 2340-2354, 2009.

[11] Jr. Dennis M. Healy, and Gustavo K. Rohde, "Fast Global Image Registration using Random Projections," 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2007, pp. 476-479, 2007.

[12] C. Fookes and A. Maeder, "Quadrature-Based Image Registration Method using Mutual Information," IEEE International Symposium on Biomedical Imaging: Nano to Macro, vol. 1, pp. 728-731, 2004.

[13] M. Petrou and P. Bosdogianni, Image Processing—The Fundamentals. New York: Wiley, 1999.

[14] Pere Marti-Puig, "A Family of Fast Walsh Hadamard Algorithms With Identical Sparse Matrix Factorization," IEEE Transactions on Signal Processing Letters, vol. 13, no. 11, pp. 672-675, 2006.

[15] J. L. Moigne, W. J. Campbell, and R. F. Cromp, "An automated parallel image registration technique based on correlation of wavelet features," IEEE Trans. Geosci. Remote Sens., vol. 40, no. 8, pp. 1849–1864, Aug. 2002.

[16] J. P. W. Pluim, J. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," IEEE Trans. Med. Imag., vol. 19, no. 8, pp. 899–814, Aug. 2000.

[17] Z. Zhang, J. Zhang, M. Liao, and L. Zhang, "Automatic registration of multi-source imagery based on global image matching," Photogramm. Eng. Remote Sens., vol. 66, no. 5, pp. 625–629, May 2000.

[18] M. Bossert, E. M. Gabidulin, and P. Lusina, "Space-time codes based on Hadamard matrices proceedings," in Proc. IEEE Int. Symp. Information Theory, Jun. 25–30, 2000, p. 283.

[19] L. Ping, W. K. Leung, and K. Y. Wu, "Low-rate turbo-Hadamard codes," IEEE Trans. Inf. Theory, vol. 49, no. 12, pp. 3213–3224, Dec. 2003.

**D.Sasikala** is presently working as Assistant Professor, Department of CSE, Bannari Amman Institute of Technology, Sathyamangalam. She received B.E.(CSE) from Coimbatore Institute of Technology, Coimbatore and M.E. (CSE) from Manonmaniam Sundaranar University, Tirunelveli. She is now pursuing Phd in Image Processing. She has 11.5 years of teaching experience and has guided several UG and PG projects. She is a life member of ISTE. Her areas of interests are Image Processing, System Software, Artificial Intelligence, Compiler Design.

R. Neelaveni is presently working as a Assistant Professor, Department of EEE, PSG College of Technology, Coimbatore. She has a Bachelor's degree in ECE, a Master's degree in Applied Electronics and PhD in Biomedical Instrumentation. She has 23 years of teaching experience and has guided many UG and PG projects. Her research and teaching interests includes Applied Electronics, Analog VLSI, Computer Networks, and Biomedical Engineering. She is a Life member of Indian Society for Technical Education (ISTE). She has published several research papers in International, National Journals and Conferences.

MULTI - LEVEL INTRUSION DETECTION MODEL USING MOBILE AGENTS IN DISTRIBUTED NETWORK ENVIRONMENT

S.Ramamoorthy
Research Scholor
Sathyabama university,Chennai
mailrmoorthy@yahoo.com

Dr.V.Shanthi
Professor
St.Joseph's college of engineering,chennai
drvshanthi@yahoo.co.in

ABSTRACT

Computer security in today's networks is one of the fastest expanding areas of the computer industry. Therefore protecting resources from intruders is a difficult task that must be automated so that it is efficient and responsive. Most intrusion-detection systems currently rely on some type of centralized processing to analyze the data necessary to detect an intruder in real time. A centralized approach can be vulnerable to attack. If an intruder can disable the central detection system, then most protection is weakened. The paper presented here demonstrates that independent detection agents can be run in a distributed fashion at three levels, each operating mostly independent of the others, thereby cooperating and communicating with the help of mobile agents to provide a truly distributed detection mechanism without a single point of failure. The agents can run along with user and system applications without much consumption of system resources, and without generating much amount of network traffic during an attack.

1. INTRODUCTION

Intrusion detection means identifying any set of actions that attempt to compromise the integrity, confidentiality or availability of resource. Intrusion Detection is the art of detecting inappropriate, incorrect, or anomalous activity. Generally there are two types of intrusion detection namely misuse detection and anomaly detection. Misuse detection deals with finding out known patterns of attack like chain loop attack, denial of service attack, etc.

Intrusion Detection Systems are broadly classified into host based system, network based system and distributed system. Intrusion Detection systems that operate on a host to detect malicious activity on that host are called host-based Intrusion Detection systems and Intrusion Detection systems that operate on network data flows are called network-based Intrusion Detection systems. The third category is the distributed intrusion detection system where IDS modules are installed on each machine and processed independently. The goal of IDS is to reduce the number of false positives as much as

possible. There are two types of intrusion detection namely, anomaly detection and misuse detection. Misuse detection deals with identifying known patterns of attacks like chain loop attack, denial of service attack, etc. while anomaly detection deals with identifying the deviation of a user from normal.

2. LITERATURE SURVEY

DIDMA [1] is a system developed to detect intrusion activities throughout the network. This system uses mobile agents that can move from one node to another within a network, and perform the task of aggregation and correlation of the intrusion related data. Here the system has static agents in all hosts which inform the mobile agent about the status of the system. The mobile agent, which roam about the network collects the data and goes to the mobile agent dispatcher. The mobile agent dispatcher dispatches the appropriate mobile agent and sends it to the victim host for processing

.AID [2] is a client-server architecture that consists of agents residing on network hosts and a central monitoring station. Information is collected by the agents and sent to the central monitor for processing and analysis. It currently has implemented 100 rules and can detect ten attack scenarios. The prototype monitor is capable of handling eight agents.

This system currently runs only on UNIX-based systems.The AAFID architecture [3] appears the most similar to the proposed work. AAFID is designed as a hierarchy of components with agents at the

lowest level of the tree performing the most basic functions. The agents can be added, started, or stopped, depending on the needs of the system. AAFID agents detect basic operations and report to a transceiver, which performs some basic analysis on the data and sends commands to the agents. A transceiver may transmit data to a transceiver on another host. If any interesting activity takes place, it is reported up the hierarchy to a monitor. The monitor analyzes the data of many transceivers to detect intrusions in the network. A monitor may report information to a higher-level monitor. The AAFID monitors still provide a central failure point in the system. AAFID has been developed into two prototypes: AAFID, which had many hard-coded variables and used UDP as the inter-host communication, and AAFID2, which was developed completely in PERL and is more robust. They run only on Unix-based systems.

In [4] a system has been presented that contains three levels each monitoring independently thereby cooperating and communicating among them selves with the help of mobile agents thus forming a distributed detection mechanism.

EMERALD [5] is a system developed by Sri International with research funding from DARPA. It is designed to monitor large distributed networks with analysis and response units called monitors. Monitors are used sparingly throughout the domain to analyze network services. The information from these monitors is passed to other

monitors that perform domain-wide correlation, obtaining a higher view of the network. These in turn report to higher-level enterprise monitors that analyze the entire network. EMERALD is a rule-based system. The target operating system has not been stated, but it is being designed as a multi-platform system. EMERALD provides a distributed architecture with no central controller or director; since the monitors are placed sparingly throughout the network, they could miss events happening on an unmonitored section. My approach is to employ agents on many hosts to attempt detection of all

suspicious activity.

3. SYSTEM ARCHITECTURE

The Intrusion Detection System is installed at three levels namely network level, subnet level and node level and the corresponding Intrusion Detection systems are called network monitor, subnet monitor and node detector.

At each level, the Intrusion Detection System includes information database, knowledge database, high level analyzing engine, log sensor module, host sensor
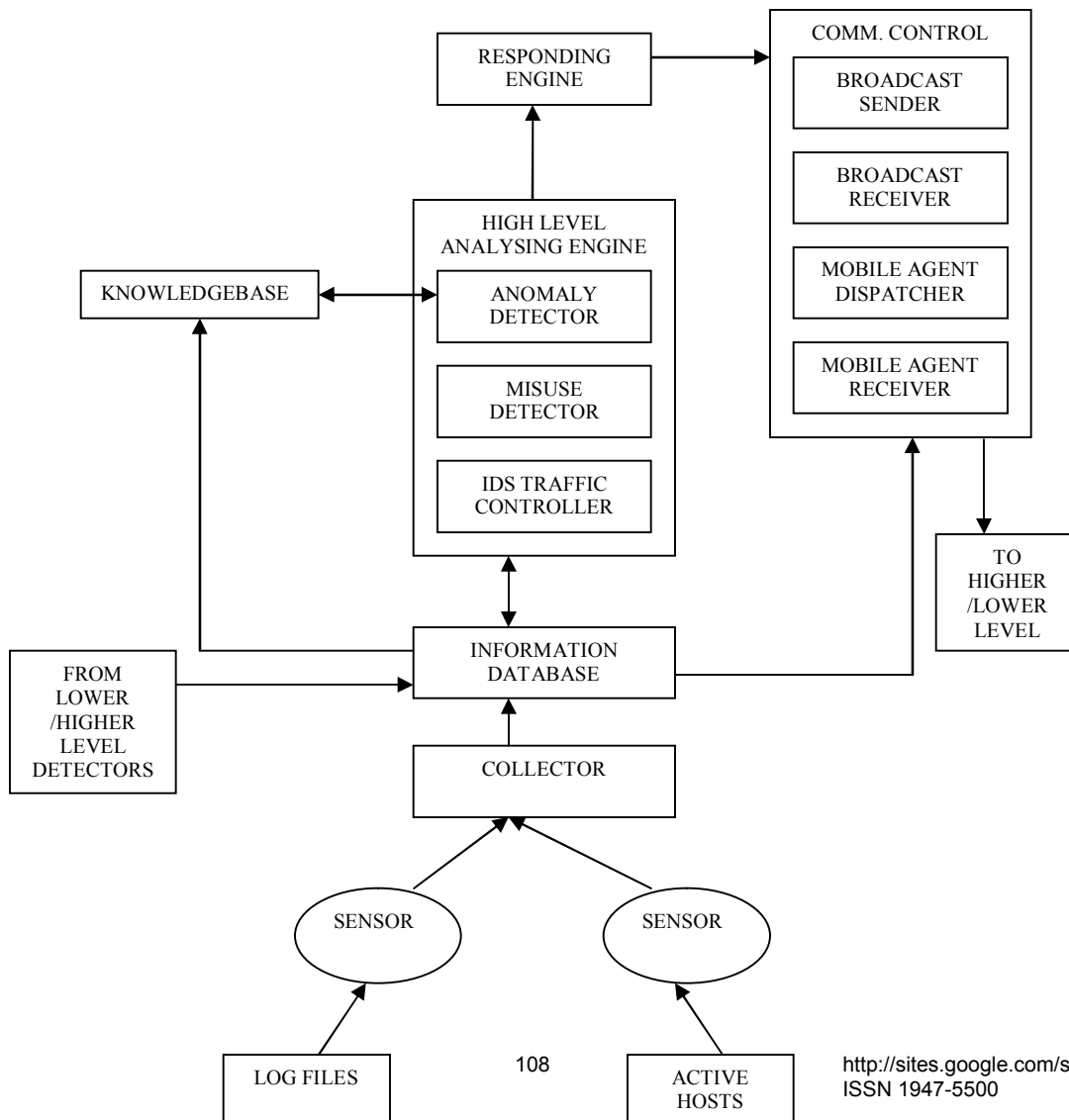
module, responding engine and communication control module.

Information database contains the data recorded from the log files and the intrusion events from the recorded host and the network systems separately. This database is accessed by the analyzing engine and mobile agents of communication control module as needed. Knowledge base contains association rules of various events for each user separately. The rules were created by using the apriori algorithm. High level analyzing engine is used to associate multi dimensional information from next level monitors and compare the information with the content of the knowledge database and judge whether intrusion has occurred or not. It is also used to find out whether any known pattern of attack is taking place. For example, if a single user tries login attempts in multiple systems each system records the attempt and is send to the high level analyzing engine and if the total number of attempts is beyond a threshold an alert containing a value is taken to the notice of the network sensor with the help of mobile agent. If the number of alerts generated by a system is noticed that it exceeds a threshold, the traffic controller of the IDS will control the traffic by stopping the messages temporarily. Communication control module helps to dispatch mobile agents as necessary. When it has found an alert in the system, this module dispatches a mobile agent which collects all the alert messages to the network sensor for further processing. Broadcast sender will broadcast its address once the

system is started to all other agents in the network. Host sensor keeps track of all the nodes and finds out whether all the nodes that registered with it at the beginning are present.

4. IMPLEMENTATION AND EVALUATION

The model being proposed can detect anomaly and misuse. The static agents for anomaly detection and misuse detection in Java 1.3 and the mobile agents have been implemented in Aglets 2.0.2. The sensors that implemented for input are Host Sensor and Log Sensor. The Host sensor works perfectly well and detects the number of hosts not responding. The Log Sensor also detects the failed login attempts and modification of protected files. Anomaly detection is also being implemented. The graph for anomaly detection is given below.

No of events	Anomalies	Detections
100	12	10
200	20	15
300	35	27
400	41	33
500	52	40
600	61	45
700	68	50
800	80	61
900	87	66
1000	95	73
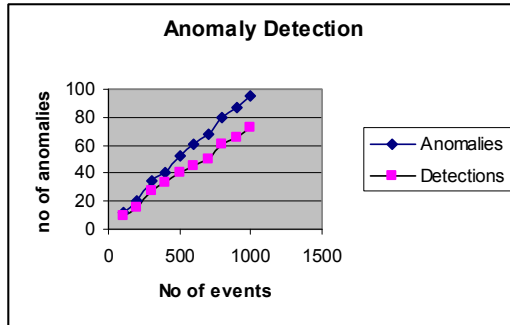
5. CONCLUSION AND FUTURE WORK

Distributed intrusion detection is considered as one of the best techniques to detect complicated attacks in high traffic flow and heterogeneous network environment. Agent

technique is good to be used in distributed environment and it provides an effective method for detecting distributed attacks. But when agent is used as a software entity it will be exposed completely to external users when it is running. So it is very important to ensure itself security of agent entity and the confidentiality and integrity of the exchanged message. Agents of IDS are deployed around network and they exchange message so as to run collaboratively. The proposed IDS are being made balanced between the agent functionality and network traffic. The autonomy and mobility of agent is fully utilized mechanism. Hence the functions of IDS are decentralized over the whole network and the single point of failure is eliminated. The proposed method is also scalable. For better detection results more sensors like file system warnings, hardware monitor, file system warnings can be added.

6. REFERENCES

[1] Pradeep Kannadiga and Mohammad Zulkernine, "DIDMA: A Distributed Intrusion Detection System Using Mobile Agents", Proceedings of the Sixth International Conference on Software Engineering, AI, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks, February 2005.

[2] Zamboni, Diego. Jai Sundar Balasubramaniyan, Jose Omar Garcia-Fernandez, David Isacoff, Eugene Spafford, "An Architecture for Intrusion Detection Using Autonomous Agents", COAST Technical Report 98/05, COAST Laboratory, Purdue University, June 1998

[3] Peter Mell, Mark McLarnon, "Mobile Agent Attack Resistant Distributed Hierarchical Intrusion Detection Systems", Proceedings of the 2nd International Workshop on Recent Advances in Intrusion Detection, West layfayette, Indiana, USA, September 7-9, 1999.

[4] Shi Zhicai, Ji Zhenzhou Hu Mingzeng, "A Novel Distributed Intrusion Detection Model Based on Mobile agent", ACM InfoSecu04, November 14-16, 2004.

[5] Neumann, Peter, G., and Phillip A. Porras, "Experience with EMERALD to Date", Proceedings 1st USENIX Workshop on Intrusion Detection and Network Monitoring Santa Clara, CA, April 1999.

[6] Wenke Lee, Stolfo, S.J., Chan, P.K., Eskin, E., Wei Fan: Miller, M, Hershkop, S.,Junxin Zhang "Real time data mining-based intrusion detection", DARF'A Information Survivability Conference & Exposition II,2001 Proceedings,

Volume: 1. 12-14 June 2001 Pages: 89 ~ 100 vol.1

[7] Tomoaki Kaneda, Youhei Tanaka, Tomoya Enokido, Makoto Takizawa, "Transactional agent model for Fault tolerant systems", ACM Symposium of Applied Computing, March 13-17, 2005.

Defending AODV Routing Protocol Against the Black Hole Attack

Fatima Ameza,
Department of computer sciences,
University of Bejaia, 06000
Algeria.

Nassima Assam,
Department of computer sciences,
University of Bejaia, 06000
Algeria.

Rachid Beghdad
Department of computer sciences,
University of Bejaia, 06000
Algeria.

Abstract—**In this paper we propose a simple method to detect Black hole attacks in the Ad hoc On Demand Vector (AODV) routing protocol. Even if many previous works focused on authentication and cryptography techniques, nevertheless these techniques suffer from some weaknesses. In fact, this kind of solution is just a first line of defense, which should be completed by an intrusion detection system as a second line.**

The second line which is proposed here consists of including the source route in the header of the control packets (RREQ). In addition to that, any intermediate node records the sequence number of the destination. Thus, if the packet is compromised, the destination node can easily retrieve the address of the attacker. To secure RREP packets, any intermediate node records the addresses of the nodes to which it forwards RREQ. Thus, any node receiving RREP can check if the sender is legitimate or not. Simulation results show the robustness of our protocol and that it allows delivering a high ratio of data and consumes less route establishment delay.

Keywords-component; AODV routing protocol; Black hole attacks; Intrusion detection; Reactive routing protocols; Wireless ad hoc networks.

I. INTRODUCTION

Wireless networks are inherently susceptible to security problems. The intrusion on the transmission medium is easier than for wired networks and it is possible to conduct denial of service attacks by scrambling the used frequency bands. The ad hoc context increases the number of potential security vulnerabilities. Because by definition without infrastructure, ad hoc networks can not benefit from the security services offered by dedicated equipment: firewalls, authentication servers, etc... The security services must be distributed, cooperative and consistent with the available bandwidth. Routing also poses specific problems: each node in the network can serve as a relay and is able to capture or divert traffic in transit. The work presented here is in this context.

We address here the problem of securing the AODV routing protocol against the Black Hole attack.

During routing in a mobile ad hoc network (MANET), if no control is done on the origin and integrity of the routing message of the network, a malicious node can easily cause disturbances. This will be even easier than wireless ad hoc networks have no physical barrier to protect themselves and all elements can potentially participate in the routing mechanism. If a malicious node has the ability to compromise a valid network node, it can at the discovery process respond to route initiator node with a *route reply* message by announcing a minimal cost path, to the target node. The transmitter node will then update its routing table with the wrong information. The data packet of the transmitter node will be relayed to the target node by the malicious node that can simply ignore them. This attack is called a "black hole". The packets are picked up and absorbed by the malicious node. This is an example of attack that may occur in a wireless ad hoc network routing protocol.

The first approach of securing the AODV protocol has been made by Zapata with his Secured AODV (SAODV) [1]. In a second publication [2] the protocol is presented in greater detail. SAODV which is based on public key cryptography extends the AODV message format to include security parameter for security the routing messages.

Adaptive Secure AODV (A-SAODV) [3] is a prototype implementation of SAODV, based on the AODV-UU implementation by Uppsala University. Unlike AODV-UU, A-SAODV is a multithreaded application: cryptographic operations are performed by a dedicated thread to avoid blocking the processing of other messages.

SecAODV [4] is a secure routing protocol, its implementation is similar to that of Boostrapping Security Associations for Routing in Mobile Ad hoc Networks (BSAR) [5] and Secure Bootstrapping and Routing in an IPv6-based ad hoc network (SBRP) [6] for DSR. SecAODV is a distributed algorithm designed for MANETs under IPv6, it did not require a trust relationship established between pairs of nodes, or synchronization between nodes, or shared key or other secure association between nodes.

M. Al-Shurman et al. [7] propose two solutions to the Black Hole attack. In the first solution the transmitter is required to authenticate the node that sent the route reply packet (RREP).

The idea here is to wait the arrival of the RREP packet from more than one node, until the identification of a safe route. In the second solution, each packet in the network must have a unique sequence number; and the following packet must have a sequence number greater than the one of the current packet. Each node records the sequence number of the packet and uses it to check if the received packet is sent by the same node or not.

C. Tseng et al [8] propose a solution based on the specification of intrusion detection to detect attacks on AODV [9], their approach is to model the behavior of AODV by a machine of finite-state (finite state machine) to detect violations of the protocol specification.

In this article we present an approach for defending AODV protocol against Black Hole attacks. Our main first idea is to include the source route in the header of the RREQ control packets. In addition to that, any intermediate node records the sequence number of the destination. Thus, if the packet is compromised, the destination node can easily retrieve the address of the attacker. On the other hand, each node forwarding a RREQ packet records the addresses of its successors in a local table. Thus, it can check if the sender of the RREP received packet is legitimate or not.

The remainder of the paper is organized as follows: Section 2 presents briefly the AODV protocol. Attacks against AODV are described in Section 3. We especially detail the Balck hole attack in this section. Our approach is described in details in section 4. Section 5 presents simulation results. Finally, section 6 concludes the paper.

II. THE AODV PROTOCOL

AODV (Ad-hoc On-demand Distance Vector) [10] is a loop-free routing protocol for ad-hoc networks. It is designed to be self-starting in an environment of mobile nodes, withstanding a variety of network behaviors such as node mobility, link failures and packet losses.

At each node, AODV maintains a routing table. The routing table entry for a destination contains three essential fields: a next hop node, a sequence number and a hop count. All packets destined to the destination are sent to the next hop node. The sequence number acts as a form of time-stamping, and is a measure of the freshness of a route. The hop count represents the current distance to the destination node.

In AODV, nodes discover routes in request-response cycles. A node requests a route to a destination by broadcasting an RREQ message to all its neighbors. When a node receives an RREQ message but does not have a route to the requested destination, it in turn broadcasts the RREQ message. Also, it remembers a *reverse-route* to the requesting node which can be used to forward subsequent responses to this RREQ. This process repeats until the RREQ reaches a node that has a valid route to the destination. This node (which can be the destination itself) responds with an RREP message. This RREP is unicast along the reverse-routes of the intermediate nodes until it reaches the original requesting node. Thus, at the end of this request-response cycle a *bidirectional* route is established between the requesting node and the destination. When a node loses connectivity to its next

hop, the node invalidates its route by sending an RERR to all nodes that potentially received its RREP. On receipt of the three AODV messages: RREQ, RREP and RERR, the nodes update the next hop, sequence number and the hop counts of their routes in such a way as to satisfy the partial order constraint mentioned above.

III. ATTACKS AGAINST AODV

Attacks against AODV can be classified in two classes [11]:

- *Passive attacks*: In a passive attack, the attacker does not disturb the routing process but only attempts to discover valuable information by listening to the routing traffic. The major advantage for the attacker in passive attacks is that in a wireless environment the attack is usually impossible to detect. This also makes defending against such attacks difficult. Furthermore, routing information can reveal relationships between nodes or disclose their IP addresses. If a route to a particular node is requested more often than to other nodes, the attacker might expect that the node is important for the functioning of the network, and disabling it could bring the entire network down.

- *Active attacks*: These attacks involve actions performed by adversaries, for instance the replication, modification and deletion of exchanged data. The goal may be to attract packets destined to other nodes to the attacker for analysis or just to disable the network. A major difference in comparison with passive attacks is that an active attack can sometimes be detected.

The following is a list of some types of active attacks that can usually be easily performed against AODV protocol.
Black hole: In the black hole attack [12], a malicious node uses the routing protocol to advertise itself as having the shortest path to the node whose packets it wants to intercept.
Black hole attack against RREQ packets: As it was said before (section 2), the sequence number of a packet acts as a form of time-stamping, and is a measure of the freshness of a route. Indeed, the node having the higher sequence number to reach a given destination node D, will be considered as the one having the shorter route to D. So, on receipt of the RREQ packet, the attacker will simply set the sequence number to the higher possible value. In this case, this malicious device will be able to insert itself between the communicating nodes, and will be able to do anything with the packets passing between them.
Black hole attack against RREP packets: Similarly, on receipt of a RREP from the legitimate destination node D, the malicious node M will set the sequence number of this packet to the higher possible value. Consequently, all the intermediate nodes between M and the source node, will forward the message of the malicious node.
Wormhole: In the wormhole attack [13], an attacker records packets (or bits) at one location in the network, tunnels them to another location, and retransmits them there into the network. The wormhole attack is possible even if the attacker has not compromised any hosts and even if all communication provides authenticity and confidentiality. The wormhole attack can form a serious threat in wireless networks,

especially against many ad hoc network routing protocols and location-based wireless security systems.

Rushing attack: This kind of attack [13] is a malicious attack that is targeted against on-demand routing protocols that use duplicate suppression at each node, like AODV. An attacker disseminates RREQs quickly throughout the network, suppressing any later legitimate RREQs when nodes drop them due to the duplicate suppression. Thus the protocol can not set up a route to the desirable destination.

Spoofing: By masquerading as another node, a malicious node can launch many attacks in a network. This is commonly known as spoofing [14]. Spoofing occurs when a node misrepresents its identity in the network, such as by altering its MAC or IP address in outgoing packets. Spoofing combined with packet modification is really a dangerous attack.

Routing table overflow: In a routing table overflow attack the attacker attempts to create routes to nonexistent nodes [15]. The goal is to create enough routes to prevent new routes from being created or to overwhelm the protocol implementation. Proactive routing algorithms attempt to discover routing information even before it is needed while a reactive algorithm creates a route only once it is needed. This property appears to make proactive algorithms more vulnerable to table overflow attacks. An attacker can simply send excessive route advertisements to the routers in a network.

Reactive protocols, on the other hand, do not collect routing data in advance. For example in AODV, two or more malicious nodes would need to cooperate to create false data efficiently. The other node requests routes and the other one replies with forged addresses.

IV. OUR APPROACH

We called our approach AODV-SABH (AODV Secured Against Black Hole attack). This is why our approach leads to secure both the RREQ and the RREP packets.

Securing RREQ packets: To secure RREQ packets we propose to add two fields in the RREQ packet. The first field will be used to include the list of the addresses of all the intermediate nodes between the source and the destination, in order to detect the address of the attacker. On the other hand, each node will use the second field to record the sequence number of the destination node that it knows. On receipt of the RREQ packet, the destination node D compares its own sequence number (SN_D) to the one of the received packet. If the sequence number of the received packet is greater than SN_D then the packet will be rejected, D will use the first added field in the packet to find the intruder, and it will alert the other nodes.

For example, the following graph (figure 1) represents a network where the node A requests a route to node D. It sends a RREQ packet having a sequence number equal to 30. On receipt of this packet, the malicious node M will set the sequence number to 1000. On receipt of the packet of node A, node B will set the sequence number to 60. Finally, the destination node D will focus on the message of M thinking that this node has the freshness route to the source node A. D will then send a RREP message to A via the node M.
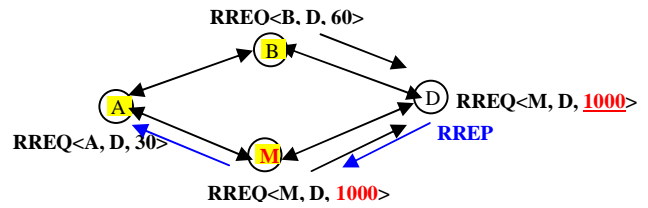
Fig. 1. Example of Black hole attack on RREQ packets.

By using AODV-SABH the node D will detect that node M is malicious, it will reject its packet and will send a RREP packet to the source node A via the legitimate node B (see figure 2). In fact, SN_D is really equal to 60, but the sequence number of the packet of M is equal to 1000 (!)
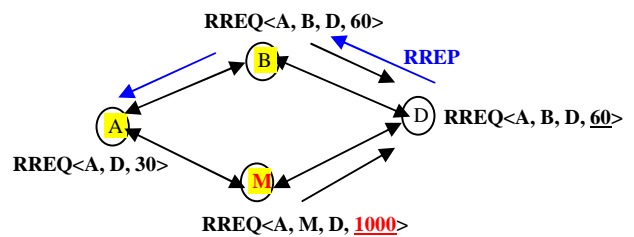
Fig. 2. Using AODV-SABH to detect the malicious node.

Securing RREP packets: To secure RREP packets, every node will record the addresses of all nodes to whom it will forward the RREQ packet in a local table. To do that, every node receiving RREQ packet during the route discovery process must sends its address to the sender. So, when a node receives a RREP packet it can check if the address of the sender belongs or not to its local table. If the address of the sender of RREP does not match any address recorded in its local table, then the receiving node concludes that the sender is a malicious node. So, it will reject the packet, and will alert the other nodes.

V. SIMULATIONS

A. Simulation parameters

For our simulations we used the Network Simulator 2 (ns-2). Our simulations consist of 20 nodes evolving in a region of (950 m × 950 m) during 100 seconds. Transmission range is set to 250 meters. Random waypoint movement model is used and maximum movement speed is 12m/s.

Packets among the nodes are transmitted with constant bit rate (CBR) of one packet per second, and the size of each packet is 512 bytes.

In these simulations we used the following evaluation metrics:

Packet delivery ratio (PDR): The percentage of data packets delivered to destination with respect to the number of packets sent. This metric shows the reliability of data packet delivery.

Control traffic: This metric informs us about the amount of control packets generated by the protocol for the research, the establishment and the maintenance of routes.

Route establishment delay (RED): This parameter shows us the time needed for the creation of a route by a source node, it is computed in milliseconds.

B. Simulation results

All the results described here are mean values of 50 experiments. Firstly, the aim of our simulation is to study the effect of the black hole attack on both the AODV and AODV-SABH protocols. This is why; by varying the number of source nodes from 10 to 15, this first experiment aims to show the impact of this parameter on the PDR. The following graph illustrates the results.
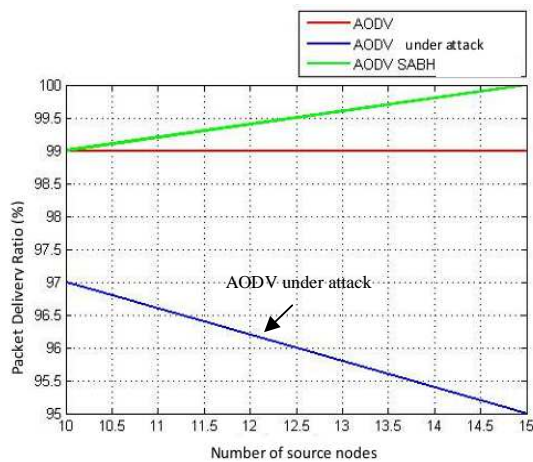
Fig. 3. The impact of the number of source nodes on the PDR.

According to figure 3, we can conclude that AODV-SABH outperforms AODV protocol in term of PDR. This is mainly due to the fact that our protocol detects the attacker and allows the source nodes to avoid it. By avoiding the attacker, our protocol finds shortest paths, and so, delivers more packets. On the other hand, the PDR decreases in the case of AODV that is subject to an attack. This is due to the fact that the number of correctly received packet is very less then the number of transmitted packets. Indeed, with the increase of the source nodes, the probability of intrusion increases, and the malicious node absorbs all the data packets passing through it.

In the following experiment we will look for the impact of the nodes mobility on the PDR, in case of AODV and AODV-SABH. We will vary the movement speed of nodes from 8 to 12 m/s and we will use 5 source nodes.
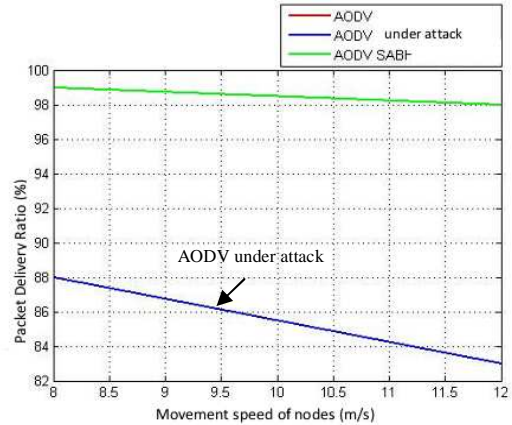
Fig. 4. The impact of the nodes mobility on the PDR.

According to figure 4, we can conclude that AODV-SABH outperforms AODV under attack in term of PDR while varying the movement speed of nodes. Even if AODV-SABH detects attackers and rejects compromised control packets; it behaves like a safe AODV (it performs the same PDR values as those of AODV). In this case, the PDR decreases lightly (from 99% to 98%) when the speed increases. In fact, when the speed increases, links between nodes may break and the source nodes must re-run the discovery route process to establish new routes. In this case, there will be more control packets transmitted and less data packets.

The PDR of AODV which is subject to an attack decreases when the movement speed of nodes increases. This is justified by the fact that when the mobility of nodes increases the network topology changes frequently, and hence the links are broken, forcing source nodes to re-run the route discovery process. Consequently, the attacker can easily exploit these new phases of route discovery to insert itself between legitimate nodes and do anything with the received packets.

In the next experiment we want to compute the cost of route discovery, while using 5 source nodes, by computing the number of control packets needed to establish a route. To do this, we computed the number of control packets (RREQ/RREP) according to the movement speed of nodes and the number of malicious nodes (from 1 to 9) in the network.
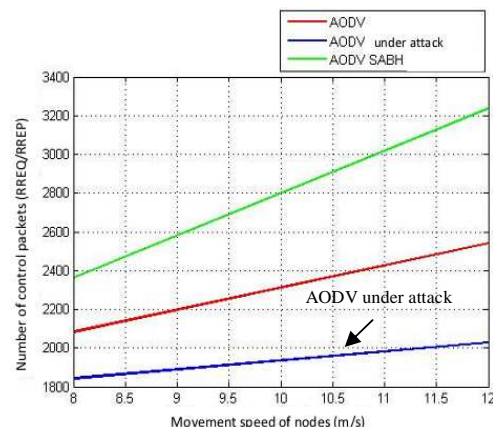
Fig. 5. The impact of the nodes mobility on the number of control packets (RREQ/RREP).

According to figure 5, the number of control packets increases whatever is the used protocol. The attacked AODV (green graph) performs the less number of control packets (RREQ/RREP). In fact, in the absence of any countermeasure against the attacker, all the source nodes believe that their established routes are correct, and do not re-run the route discovery process.

If there is no attack against AODV (red graph) we observe that the number of control packets grows with the growing of the movement speed of nodes. As said previously, this is due to the fact that links between nodes may break and the nodes must re-run the discovery route process to establish new routes.

AODV-SABH performs the higher number of control packets. Indeed, whenever the attacker is detected, this protocol re-runs the discovery route process, and rejects any compromised RREQ or RREP packets. On the other hand, there are 5 source nodes, so there are more control packets to manage. In addition to that, the nodes are moving, so, the risk of broken links increases, and then the source nodes must restart the route discovery process.
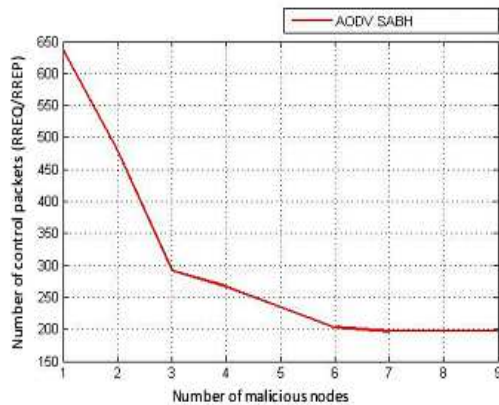
Fig. 6. The impact of the number of malicious nodes on the number of control packets (RREQ/RREP).

According to figure 6, the number of control packets decreases when the number of malicious nodes increases in case of AODV-SABH. This can be explained by the fact that our protocol detects the intruders and does not transmit any RREP packet if the received RREQ is compromised. We can also conclude that if 6 nodes among the 20 composing the network are malicious, they can compromise the whole network and our protocol is not efficient in this case. In this case the source nodes believe that their established routes are correct and do not request new routes.

Finally, the following experiment will show the impact of the number of nodes on the RED.
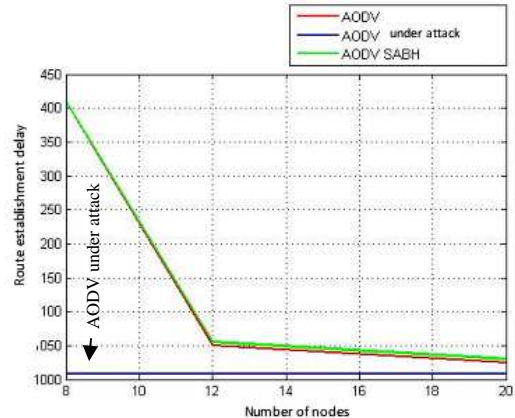
Fig. 7. The impact of the number of nodes on the route establishment delay.

Figure 7 shows that AODV-SABH behaves like AODV (without attack). Indeed, the two protocols reach the same RED values while varying the number of nodes. When the number of nodes increases, the nodes are more close to each other, and less is the delay of route establishment.

In case of the attacked AODV (without any countermeasure), the delay is constant even if the number of nodes increases. This is mainly due to the fact that the whole network is compromised and source nodes do not request new routes.

VI. CONCLUSION

An efficient and simple approach for defending the AODV protocol against Black Hole attacks is proposed. Our main contribution consists of including the source route in the header of the control messages. Indeed, each intermediate node receiving a RREQ packet adds its own address to the message. After that, it sends it to its successors. In addition to that, any node must include in such a packet the sequence number of the destination. Thus, when the destination node receives the RREQ packet, it checks if its sequence number is less than the one included in the packet. If it is, it will conclude to an attack and can find the address of the intruder by consulting the list of addresses in the RREQ packet. On the other hand, to secure RREP packets, every node sending RREQ must record the addresses of its receptors in a local table. So, when it receives a RREP packet it can check if the address of the sender is included or not in this table. Any compromised packets will be rejected and the detecting node alerts the other nodes in the network. In this case, source nodes must request new routes to reach the destination.

As future work we will focus on securing AODV against other known attacks. We will also focus on resolving the problem of multiple attacks against AODV. After that we will implement robust software to detect and counter any intruder.

REFERENCES

[1] M. G. Zapata, "Secure ad-hoc on-demand distance vector (saodv) routing,"ftp://manet.itd.nrl.navy.mil/pub/manet/2001-10.mail,October 2001.

[2] M. G. Zapata and N. Asokan, "Securing ad-hoc routing protocols," in Proceedings of the 2002 ACM Workshop on Wireless Security, pp. 1–10, Sept 2002.

[3] Davide Cerri and Alessandro Ghioni, "Securing AODV: The A-SAODV Secure Routing Prototype", IEEE Communications Magazine, Vol. 42(2), pp. 120-125, 2008.

[4] A.J. Michaele, I.Karygiannis,T Anand and al. "Secure Routing and intrusion Detection in Ad Hoc Networks", in the Proceedings of the 3rd International Conference on pervasive computing and communications(Percom 2005), Kauai Island, Hawaii. 2005.

[5] V.R.G.Bobba, L.Eschenauer and W.Arbaugh. Bootstarpping Security Association for Routing in Mobile Ad HocNetworks, in the Proceedings of GlobeCom'2003, pp. 1511-1515, 2003.

[6] J.R.Jiang, Y.C.Tseng and J.H.Lee. Secure Bootstrapping and routing in an IPv6-based Ad Hoc Network, ICCP Workshop on Wirless Security and Privacy, pp.375-390, 2003.

[7] M. Al-Shurman and al., "Black Hole Attack in Mobile Ad hoc Networks", in the Proceedings of ACMSE'04, pp. 96-97, 2004.

[8] C. Tseng."A Specification-based Intrusion Detection System for AODV", in the Proceeding of the lst ACM Workshop Security of Ad Hoc and Sensor Networks Fairfax, pp. 125-134, 2003.

[9] E.M.Beldin Adg-Royer, C.E.Perkins and S.Das. "Ad hoc on demand distance vector (aodv) Routing", IETF Internet draft, draft-ietf-manet-aodv-12.txt, 2002.

[10] Madanlal Musuvathi, David Y. W. Park, Andy Chou, Dawson R. Engler, David L. Dill: "CMC: A Pragmatic Approach to Model Checking Real Code". In the Proceedings of OSDI' 2002, pp. 75-88, 2002.

[11] Qifeng Lu , " Vulnerability of Wireless Routing Protocols ", internal report, University of Massachusetts Amherst, Dec 15, 2002.

[12] Feiyi Wang, Brian Vetter and Shyhtsun Wu. Secure Routing Protocols: Theory and Practice. North Carolina State University, May 1997.

[13] Y.-C. Hu, A. Perrig, and D. B. Johnson. Ariadne: A secure on-demand routing protocol for ad hoc networks. In Proceedings of the 8th ACM International Conference on Mobile Computing and Networking. (MobiCom), pp. 21-38, 2002.

[14] K. Sanzgiri, B. Dahill, B. N. Levine, C. Shields, and E. M. Belding-Royer. A secure routing protocol for ad hoc networks. In Proceedings of the 10th IEEE InternationalConference on Network Protocols (ICNP), pp. 78-87, 2002.

[15] www.tcm.hut.fi/Opinnot/Tik-110.501/2000/papers/lundberg.ps

AUTHORS PROFILE

Fatima Ameza obtained Master degree in computer sciences from the University of Bejaia in 2009. She is currently a PhD student in the RESYD doctoral school of Bejaia university. His research topic focuses on securing wireless networks.

Nassima Assam obtained Master degree in computer sciences from the University of Bejaia in 2009.

Rachid Beghdad, received his computer science engineer degree in 1991 from the ENITA school of engineers, Algiers, Algeria. He received his Master computer science degree from Clermont-Ferrand University, France, in 1994. He earned his Ph.D. computer science degree from Toulouse University, France, in 1997. He obtained his Habilitation from the University of Constantine, 2010.

He is a reviewer for some journals, such as the *Advances in Engineering Software* journal, Elsevier, UK, the *Computer Communications* journal, Elsevier, UK, the WESEAS transactions on computer journal, Greece, and the IJCSSE journal, UK. He was also a reviewer for the CCCT'04, CCCT'05, CCCT'09, and CCCT'10 International Conferences, USA.

His main current interest is in the area of computer communication systems including intrusion detection methods, wireless sensor networks, unicast and multicast routing protocols, real-time protocols, and wireless LAN protocols.

An Efficient OFDM Transceiver Design suitable to IEEE 802.11a WLAN standard

T.Suresh

Research Scholar, R.M.K Engineering College
Anna University, Chennai
TamilNadu, India
fiosuresh@yahoo.co.in

Dr.K.L.Shunmuganathan

Professor & Head, Department of CSE
R.M.K Engineering College, Kavaraipettai
TamilNadu, India
kls_nathan@yahoo.com

Abstract—**In today's advanced Communication technology one of the multicarrier modulations like Orthogonal Frequency Division Multiplexing (OFDM) has become broadened, mostly in the field of wireless and wired communications such as digital audio/video broadcast (DAB/DVB), wireless LAN (802.11a and HiperLAN2), and broadband wireless (802.16). In this paper we discuss an efficient design technique of OFDM transceiver according to the IEEE 802.11a WLAN standard. The various blocks of OFDM transceiver is simulated using ModelSimSE v6.5 and implemented in FPGA Xilinx Spartan-3E Platform. Efficient techniques like pipelining and strength reduction techniques are utilized to improve the performance of the system. This implementation results show that there is a remarkable savings in consumed power and silicon area. Moreover, the design has encouraged the reduction in hardware resources by utilizing the efficient reconfigurable modules.**

Keywords-**FPGA**; **VHDL**; **OFDM**; **FFT**; **IFFT; IEEE 802.11a**

I. INTRODUCTION

Wireless communications are evolving towards the Multi-standard systems and other communication technologies, are utilizing the widely adopted Orthogonal Frequency Division Multiplexing (OFDM) technique, among the standards like IEEE 802.11a&g for Wireless Local Area Networks (WLANs), Wi-Fi, and the growing IEEE802.16 for Metropolitan Access, Worldwide Interoperability for Microwave Access (WIMAX)[1]. The fast growth of these standards has helped the way for OFDM to be among the widely adopted standards and to be the fundamental methods for the improvements of the next generation telecommunication networks. In broadband wireless communication, designers need to meet a number of critical requirements, such as processing speed, flexibility, and fast time to market. These requirements influence the designers in selecting both the targeted hardware platform and the design

tool. Therefore, to support high data rates and computational intensive operations, the underlying hardware platform must have significant processing capabilities. FPGAs, here, promotes itself as a remarkable solution for developing wireless LAN (802.11a and HiperLAN2), and broadband wireless systems (802.16) with their computational capabilities, flexibility and faster design cycle[2]. Therefore, to support high data rates and computational intensive operations, the underlying hardware platform must have significant processing capabilities. The aim of this paper is to implement the reconfigurable architecture for the digital baseband part of an OFDM transceiver that conforms the 802.11a standard, by including 16 QAM modulator, FFT (Fast Fourier Transform) and IFFT (Inverse Fast Fourier Transform), serial to parallel and parallel to serial converter using hardware programming language VHDL (VHSIC Hardware Description Language). Moreover, this design is area and power efficient by making the use of strength reduction transformation technique that will reduce the number of multipliers used to perform the computation of FFT/IFFT processing.

The paper is organized as follows: Section II describes the OFDM point to point system. Section III represents the simulated methods of OFDM blocks and their results. Section IV briefs about the pipelining process. Section V explains the FFT/IFFT implementation by using Strength Reduction technique. Section VI shows the implementation results and resource reductions. Section VII concludes the paper.
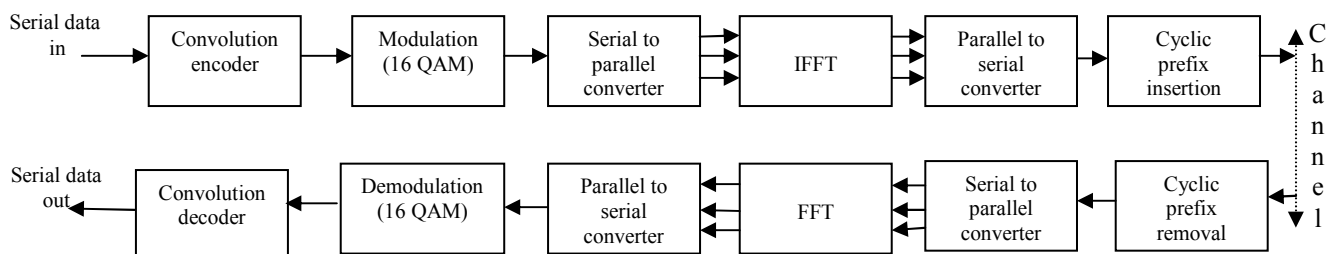
Figure 1.   OFDM point to point System

II. OFDM POINT TO POINT SYSTEM

The simplest form of a point-to-point OFDM system could be considered as transmitter building blocks into the receiver side. It represents the basic building blocks that are used in both the transmission and reception sides as shown in Fig. 1.

A. Convolution Encoder

Convolution encoder is used to create redundancy for the purpose of secured transmission of data. This helps the system to recover from bit errors during the decoding process. The 802.11a standard recommends to producing two output bits for each input. To achieve higher data rates, some of the redundant bits are removed after the encoding process is completed.

B. QAM Modulation

QAM (Quadrature Amplitude Modulation) is widely used in many digital radio and data communications. It also considers the mixture of both amplitude and phase modulation. In this paper we used 16 bit QAM and is used to refer the number of points in constellation mapping. This is because of QAM achieves a greater distance between adjacent points in the I/Q plane by distributing the points more evenly. By this way the points in the constellation are distinct and due to this, data errors are reduced.

C. IFFT/FFT

The key kernel in an OFDM transceiver is the IFFT/FFT processor. In WLAN standards it works with 64 carriers at a sampling rate of 20 MHz, so a 64-point IFFT/FFT processor is required. The Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT) are derived from the main function which is called Discrete Fourier Transform (DFT). The idea of using FFT/IFFT instead of DFT is that the computation can be made faster where this is the main criteria for implementation. In direct computation of DFT the computation for N-point DFT will be calculated one by one for each point. But for FFT/IFFT, the computation is done simultaneously and this method helps to save lot of time, and so this is similar to pipelining method[4].

The derivation starts from the fundamental DFT equation for an N point FFT. The equation of IFFT is given as shown in (1) and the equation of FFT is given as shown in (2)

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W^{-nk}, \quad k = 0,1,\dots N-1 \qquad (1)$$

$$x(n) = \sum_{k=0}^{N-1} X(k) W^{nk}, \quad k = 0,1,\dots N-1 \qquad (2)$$

where the quantity W_N^{nk} (called Twiddle Factor) is defined as

$$W_N^{nk} = e^{-j2\pi nk/N} \qquad (3)$$

This factor is calculated and put in a table in order to make the computation easier and can run simultaneously. The Twiddle Factor table is depending on the number of points used. During the computation of FFT, this factor does not need to be recalculated since it can refer to the Twiddle factor table, and thus it saves time.
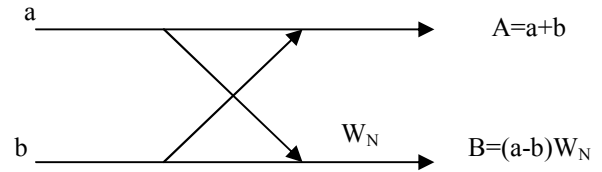
Figure 2. 2 Point Butterfly structure

D. Strength Reduction Transformation

Fig. 2 shows the 2 point Butterfly structure where multiplication is performed with the twiddle factor after subtraction. Consider the problem of computing the product of two complex numbers R and W

$$X = RW = (R_r + jR_i)(W_r + jW_i)$$
$$= (R_r W_r - R_i W_i) + j(R_r W_i + R_i W_r) \qquad (4)$$

The direct architectural implementation requires a total of four multiplications and two real additions to compute the complex product as shown in (4). However, by applying the Strength Reduction transformation we can reformulate (4) as

$$X_r = (R_r - R_i)W_i + R_f(W_r - W_i) \qquad (5)$$
$$X_i = (R_r - R_i)W_i + R_i(W_r + W_i) \qquad (6)$$

It is clearly shown as given in (5) and (6), by using the Strength Reduction transformation the total number of real multiplications is reduced to only three. This however is at the expense of having three additional adders. So in this paper the above discussed strength reduction transformation technique is used in the implementation of OFDM transceiver while multiplying the transmitted/received signal by twiddle factor.
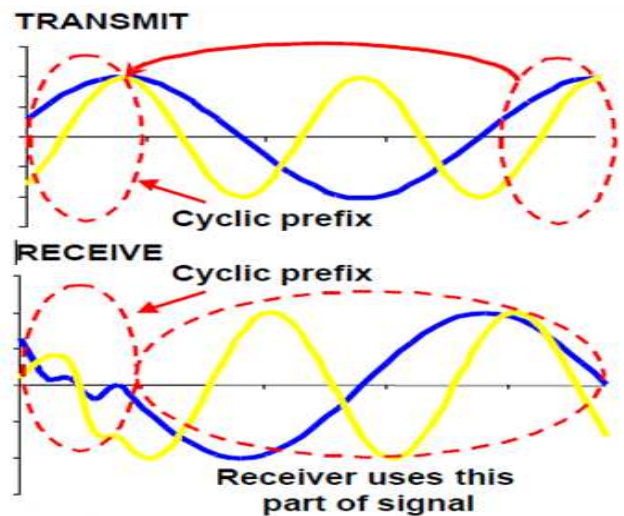
Figure 3. Cyclic Prefix

E. Cyclic Prefix

One of the most important properties of OFDM transmission is its robustness against multi path delay. This is especially important if the signal's sub-carriers are to retain their orthogonality through the transmission process. The addition of a guard period between transmitted symbols can be used to accomplish this. The guard period allows time for multipath signals from the previous symbol to dissipate before the information from the current symbol gets recorded. The most effective guard period is a cyclic prefix, which is appended at the front of every OFDM symbol. The cyclic prefix is a copy of the last part of the OFDM symbol, and is of equal or greater length than the maximum delay spread of the channel as shown in Fig. 3.

III. SIMULATED METHODS AND RESULTS

In this paper the simulated blocks of OFDM transceiver are explained and the results were analyzed. The blocks those are simulated using ModelSim SE v6.5 are given in Fig. 4. The blocks consist of OFDM transmitter which includes 16 QAM modulator and IFFT and OFDM receiver which includes FFT and 16 QAM demodulator.

In the initial stage the serial binary data value can be applied to the transmitter block through convolution encoder for the purpose of secured data transmission and modulated by the 16-QAM because of its advantageous compared to other modulations like BPSK, QPSK. An OFDM carrier signal is the sum of a number of orthogonal sub-carriers, with baseband data modulation (QAM) and it is demultiplexed into parallel streams, and each one mapped to a complex symbol stream using 16-QAM modulation.
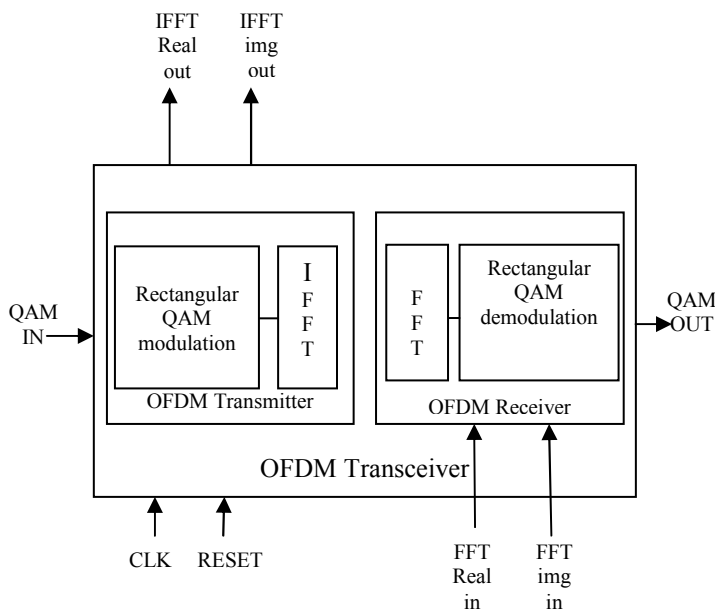
Figure 4. Simulated Blocks of OFDM Transceiver

An inverse FFT is computed on each set of symbols, delivering a set of complex symbols. The real and imaginary components (I/Q) are used to modulate the cosine and sine waves at the carrier frequency respectively, these signals are summed to give the transmitted signal. The baseband signals are sampled and passed through the OFDM receiver in FPGA and a forward FFT is used to convert back to the frequency domain. This returns of parallel streams, is converted to a binary stream using an 16-QAM demodulator. These are re-combined into a serial stream, is an estimate of the original binary stream at the transmitter. The cyclic prefix is used in OFDM Transceiver for the purpose of eliminating the ISI. This overall simulation part is done by ModelSim SE v6.5 software with VHDL language and simulated results are shown in Fig. 5.
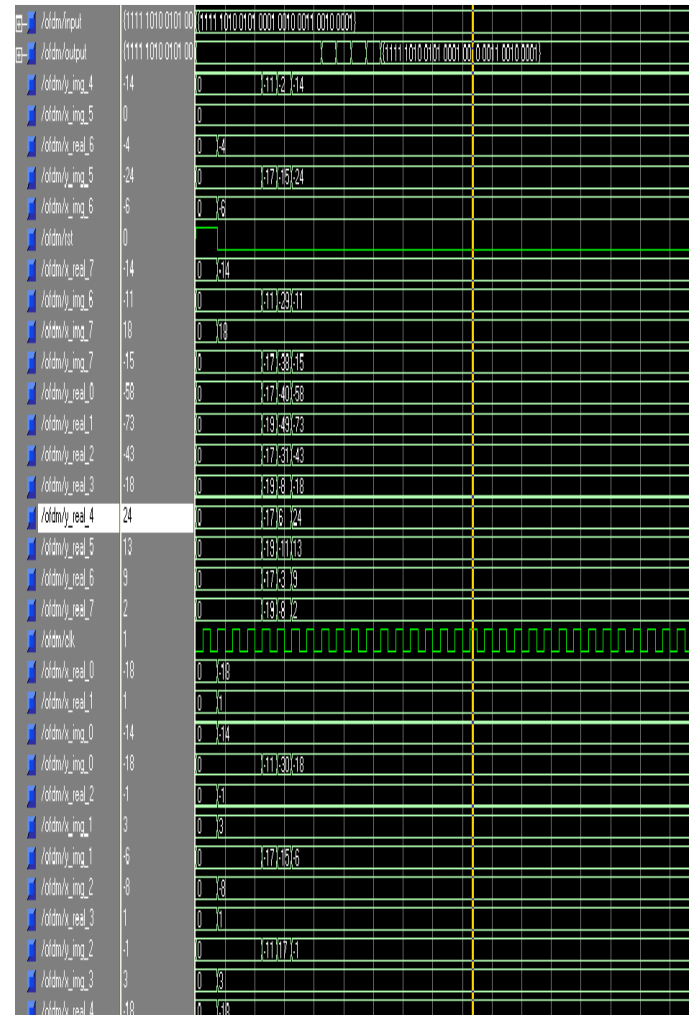
Figure 5. Simulated Results

IV. PIPELINE PROCESS

Each block in this architecture is designed and tested separately, and later those blocks are assembled and extra modules are added to compose the complete system. The design makes use of pipelining process and this is mainly achieved through duplicating the memory elements like registers or RAMs in simulation function processing and it will buffer the incoming stream of bits while the previous stream is being processed. The design environment is completely based on the Xilinx Integrated Software Environment (ISE) and
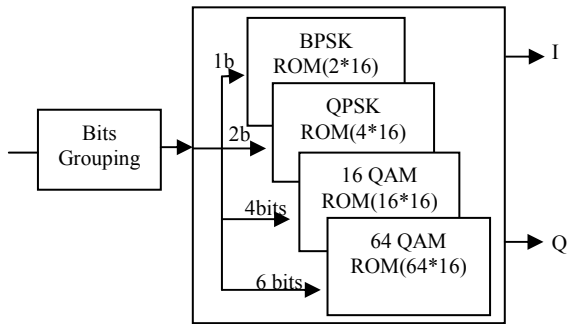
Figure 6.   Mapper Architecture

implemented in the Xilinx Spartan-3E FPGA. As a first step, the data stream is encoded using a convolution encoder, which uses a number of delay elements by representing the D-type Flip-flop for duplicating purpose. The final purpose of the coding stage is to provide the receiver with the capability to detect and correct errors through redundancy. By using this design, the need of more number of multiplexers is avoided and the abundant memory inside the FPGA is used. To perform the Pipeline process, the bits are translated or mapped into two components the In-phase and the Quadrature of (I/Q) components, those are mapped as shown in Fig. 6.

The representation of these I and Q values is based on a fixed point representation. Depending on the data rate selected, the OFDM sub-carriers are modulated using 16-QAM. This capability came from the pipelining provided by the previous and the next stages, where each generated I/Q pair is fed to the IFFT processor. The generated real and imaginary Pairs are forwarded to the Cyclic Prefix block. The last samples of the generated OFDM symbol are copied into the beginning to form the cyclic prefix. In the 802.11a standard, the last samples of the Pipelining IFFT output are replicated at the beginning to form a complete samples of OFDM symbol. These samples are considered as the maximum delay in the multipath environment.

V.   FFT/IFFT IMPLEMENTATION

FFT/IFFT computation is performed using strength reduction transformation technique in this paper. Fig. 7 shows the Processing Element(PE) and its resources used to perform FFT/IFFT computation. This implementation is compared with the direct computation of FFT/IFFT. It is demonstrated that there are four multipliers used in the direct computation of FFT/IFFT, but the number of multipliers used in the implementation of strength reduction transformation technique is reduce to only three.

VI.   IMPLEMENTATION RESULTS

The work presented in this paper is to implement the capability of an OFDM transceiver standard in a pure VHDL code implementation, and to encourage the reduction in hardware resources by utilizing the efficient techniques and suitable reconfigurable platform. The approach of divide and conquer is used to design and test each entity alone and helps to make the complete system. The work has accomplished the
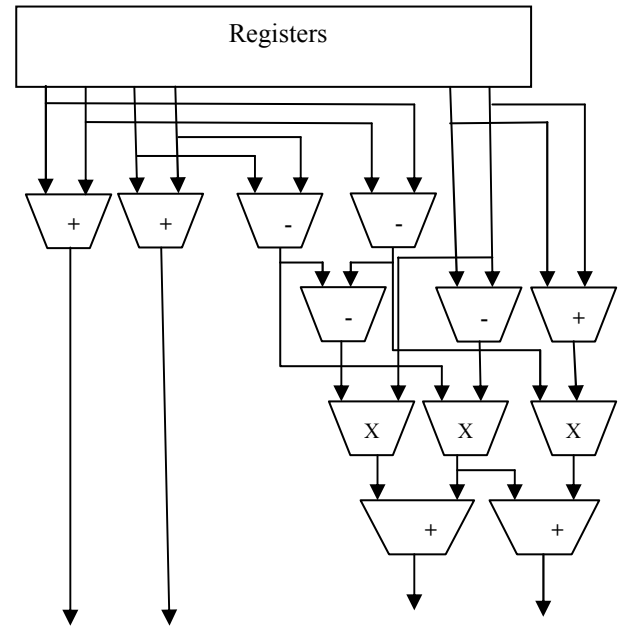
Figure 7.   PE and its resources of FFT/IFFT block

task of designing the digital baseband part of an OFDM transceiver that conforms to the IEEE 802.11a standard. However, the implemented design supports only the data rates 6, 12 and 24 Mbps in the standards.

Table I shows the resources used for implementing the blocks of OFDM system and also shows the percentage of device utilization by this design from the available resources on FPGA and the memory elements of estimated values. From this table we understood that the number of multiplexers is reduced by using the efficient pipelining and strength reduction transformation methods, and the total number of resources is also reduced remarkably.

TABLE I.   COMPLETE SYSTEM RESOURCES

Device Utilization Summary(Estimated Value)			
Logic Utilization	Used	Available	Utilization
Number of Slices	1521	3584	42%
Number of Slice Flip-Flops	1682	7168	23%
Number of 4 input LUTs	2549	7168	35%
Number of bonded IOBs	66	141	46%
Number of MULT16x16s	12	16	75%
Number of GCLKs	1	8	12%

VII.   CONCLUSION

Orthogonal Frequency Division Multiplexing is an important technology because so many developing

communication standards require OFDM because of its high throughput and multi-path. Due to this time spreading analysis and also the elimination of Inter-Symbol Interference (ISI), OFDM has several unique properties that make it especially well suited to mobile wireless data applications. In this paper the simulated and implemented results of an OFDM transceiver system through pipelining process is presented. FFT/IFFT blocks of OFDM transceiver system is implemented using strength reduction transformation method. From the result presented in this paper, it is shown that the number of hardware resources is reduced in this implementation by exploiting the efficient reconfigurable architecture. The design is implemented using a pure VHDL language in the XILINX Spartan-3E Board, and the results showed that this implementation is an efficient method in terms of Size and Resources.

REFERENCES

[1] Ahmad Sghaier, Shawki Areibi and Bob Dony, "A Pipelined Implementation of OFDM transmission on Reconfigurable Platforms", proceedings of the IEEE Conference on Communication Systems, 2008.

[2] Ahmad Sghaier, Shawki Areibi and Robert Dony "IEEE802.16-2004 OFDM Functions Implementation on FPGAS with design exploration", in Proceedings of the International Conference on Field Programmable Logic and Applications, pp. 519–522, 2008.

[3] T. Ha, S. Lee, and J. Kim. "Low-complexity Correlation System for Timing Synchronization in IEEE 802.11a Wireless LANs", In RAWCON '03: Radio and Wireless Conference, 2003. Pages 51–54, August 2003.

[4] S.B.Weinstein and P.M.Ebert, "Data Transmission by Frequency Division Multiplexing Using the Discrete Fourier Transform", IEEE Transactions on Communication Technology", Vol. COM-19, pp. 628-634, October 1971.

[5] M.Speth, S.Fechtel, G.Fock, H.Meyr, "Optimum Receiver Design for OFDM-Based Broadband Transmission-Part II: A Case Study", IEEE Transactions. On Communications, vol. 49, no. 4, pp. 571-578, April 2001.

[6] Zhi Yong Li, a thesis of "OFDM Transceiver Design with FPGA and demo on de2-70 board", July 2008.

[7] IEEE Std 802.11a-1999, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: high speed physical layer in the 5 GHZ band", July 1999.

[8] Yiyan Wu, William Y.Zou, "Orthogonal Frequency Division Multiplexing: A Multi-Carrier Modulation Scheme", IEEE Transactions on Consumer Electronics, Vol. 41, No. 3, August 1995.

[9] V.Szwarc and L.Desormeaux, "A Chip Set for Pipeline and Parallel Pipeline FFT Architectures" ,JOURNAL on VLSI Signal Processing, vol. 8, pp.253–265, 1994.

[10] K.Chang, G.Sobelman, E.Saberinia and A. Tewfik, "Transmitter Archeiticture for Pulsed OFDM" ,in the proceedings. of the 2004 IEEE Asia-Pacific conf. on circuits and systems, Vol. 2,Issue 6-9, Tainan, ROC, Dec. 2004.

[11] J.I. Smith, "A Computer Generated Multipath Fading Simulation for Mobile Radio" ,IEEE Trans. Veh. Technol., vol. VT-24, pp. 39–40, August 1975.

[12] G.Leus, S. Zhou, and G. B. Giannakis, "Orthogonal Multiple Access over Time and Frequency-selective Channels" ,IEEE Transactions on Information Theory, vol. 49, no. 8, pp. 1942–1950, August 2003.

[13] Y. G. Li and L. J. Cimini, "Bounds on the Inter Channel Interference of OFDM in time-varying impairments", IEEE Transactions on Communications, vol. 49, no. 3, pp. 401–404, March 2001.

[14] A.M.Sayeed, A.Sendonaris, and B.Aazhang, "Multiuser Detection in Fast Fading Multipath Environment" ,IEEE Journal on Selected Areas in Communications, vol. 16, no. 9, pp. 1691–1701, December 1998.

[15] S. He, M. Torkelson, "Designing Pipeline FFT Processor for OFDM De-modulation" ,in Proceedings. 1998 URSI International Symposium on Signals, Systems, and Electronics Conf., Sept. 1998.

[16] E. Bidet, D. Castelain, C. Joanblanq and P. Stenn, "A fast Single-chip Implementation of 8192 Complex Point FFT" ,IEEE J. Solid-State Circuits, March 1995.

[17] Y.Chang, K. K. Parhi, "Efficient FFT Implementation using Digit-serial arithmetic" , IEEE Workshop on Signal Processing Systems, SiPS99, 1999

[18] S.Barbarossa and A. Scaglione, "Signal Processing Advances in Wireless and Mobile Communications", Upper Saddle River (NJ), USA: Prentice-Hall, Inc., 2000, vol. 2, chap. Time-Varying Fading Channels.

[19] H Heiskala, J .T. Terry, "OFDM Wireless LANs : A Theoretical and Practical guide", Sams Publishing, 2002.

[20] M.J. Canet, F. Vicedo, V. Almenar, J. Valls, and E.R.delima. "An FPGA Based Synchronizer Architecture for Hiperlan/2 and IEEE 802.11a WLAN Systems", In PIMRC 2004: 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pages 531–535, September 2004.

[21] T. Ha, S. Lee, and J. Kim "Low-complexity Correlation System for Timing Synchronization in IEEE 802.11a Wireless LANs", in the proceedings of AWCON '03: Radio and Wireless Conference, 2003. pages 51–54, August 2003.

[22] K.Wang, J. Singh, and M. Faulkner. "FPGA Implementation of an OFDM WLAN Synchronizer", In DELTA 2004: Second IEEE International Workshop on Electronic Design, Test and Applications, 2004., pages 89–94, January 2004.

[23] T.Kim and S.C. Park. "A New Symbol Timing and Frequency Synchronization Design for OFDM-based WLAN Systems", In ICACT 07, pages 1669–1672, February 2007.

[24] F.Manavi and Y. Shayan. "Implementation of OFDM modem for the Physical Layer of the IEEE 802.11a Standard Based on Xilinx Virtex-II FPGA", In IEEE 59th Vehicular Technology Conference, 2004, pages 1768–1772, May 2004.

AUTHORS PROFILE

T.Suresh received his BE and ME degrees in Electronics and Communication Engineering from Madras University and Alagappa Chettiar College of Engineering and Technology in 1991 and 1996, respectively, and pursuing Ph.D from Anna University, Chennai ,India. Currently, he is an Assistant Professor in the Department of Electronics and Communication Engineering at R.M.K Engineering College, Chennai, India. His Research interests include FPGA Design, Reconfigurable Architecture, Multiagent System.

Dr.K.L.Shanmuganathan B.E, M.E., M.S., Ph.D working as Professor & Head, Department of Computer Science & Engg., RMK Engineering College, Chennai, TamilNadu, India. He has more than 15 publications in National and International Journals. He has more than 18 years of teaching experience and his areas of specializations are Artificial Intelligence, Networks, Multiagent Systems, DBMS.

Comparitive Analysis of Beamforming Schemes And Algorithems of Smart Antenna Array : A Review

Abhishek Rawat , R. N. Yadav and S. C. Shrivastava

Maulana Azad National Institute Of Technology
Bhopal, INDIA

Abstract— **The smart antenna array is a group of antennas in which the relative phases of the respective signals feeding the antennas are varied in such a way that the effective radiation pattern of the array**http://glossary.its.bldrdoc.gov/fs-1037/dir-003/_0364.htm **is reinforced in a desired direction and suppressed in undesired directions.** Smart antenna **are the array with smart signal processing algorithms used to identify spatial signal signature such as the direction of arriving of the signal, and use it to calculate beam forming vector, to track and locate the antenna beam on the mobile/target. An array antenna may be used to point a fixed radiation pattern, or to scan rapidly in azimuth or elevation. This paper explains the architecture; evolution of smart antenna differs from the basic format of antenna. The paper further discusses different Beamforming schemes and algorithms of smart antenna array.**

I. INTRODUCTION

In the past, wireless communication systems are deployed with fixed antenna system with fixed beam pattern. Such configuration can not meet all the requirements of modern communication environments. Smart antennas [1]-[2] are the technology that use a fix set of antenna elements in an array. The signals from these antenna elements are combined to form a movable beam pattern that can be steered to the direction of the desired user. This characteristic makes the smart antenna and minimizes the impact of noise, interference, and other effects that degrade the signal quality. The adoption of smart antenna techniques in future wireless systems is expected to have a significant impact on the efficient use of the spectrum, the minimization of the cost of establishing new wireless networks, the optimization of service quality, and realization of transparent operation across multi technology wireless networks [2]-[5]. Smart antenna systems consist of multiple antenna elements at the transmitting and/or receiving side of the communication link, whose signals are processed adaptively in order to exploit the spatial dimension of the mobile radio channel as shown in Fig.1. A smart antenna receiver can decode the data from a smart antenna transmitter this is the highest-performing configuration or it can simply provide array gain or diversity gain to the desired signals transmitted from conventional transmitters and suppress the interference. No manual placement of antennas is required.

The smart antenna electronically adapts to the environment by looking for pilot tones or beacons or by recovering certain
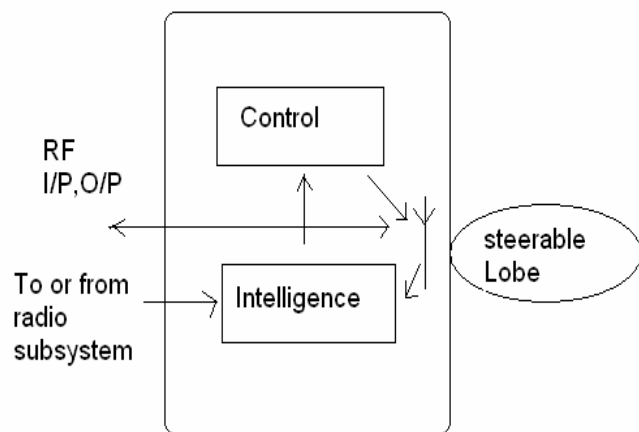
Fig. 1. Principle of smart antenna.

characteristics (such as a known alphabet or constant envelope) that the transmitted signal is known to have. The base station antennas have up till now been omni directional or sectored. This can be regarded as a "waste" of power as most of it will be radiated in other directions than toward the user and the other users will experience the power radiated in other directions as interference [4]. The idea of smart antennas is to use base station antenna patterns that are not fixed, but adapt to the current radio conditions. This can be visualized as the antenna directing a beam toward the communication partner only.

II. TYPES AND GEOMETRY OF SMART ANTENNA SYSTEMS

Smart antenna systems can improve link quality by combating the effects of multi-path propagation or constructively exploiting the different paths, and increase capacity by mitigating interference and allowing transmission

of different data streams from different antennas [6]. Smart antenna system technologies include intelligent antennas,
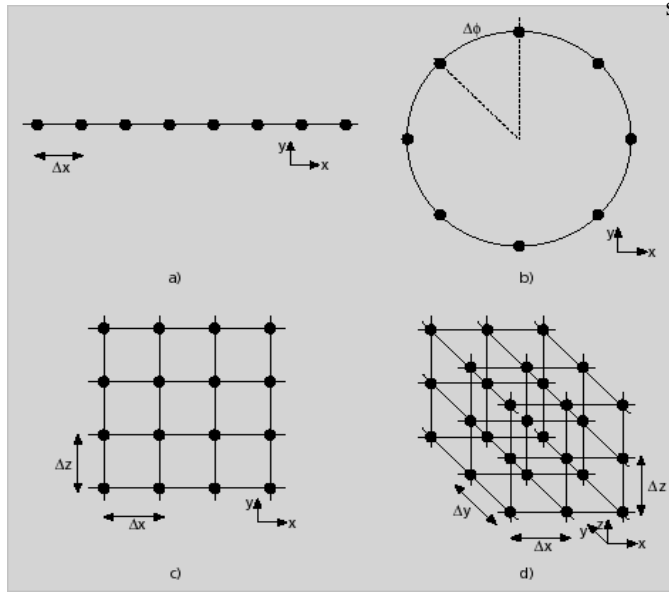
Fig.2 Different array geometries for smart antennas
a) Uniform linear array   b) Circular array;
c) 2-Dimensional grid array   d) 3-Dimensional grid array

TABLE I.   COMPARISON BETWEEN THREE BASIC TYPE OF SMART ANTENNA.

S. No	Switched Lobe	Dynamically Phased Array	Adaptive Array
1.	A finite number of fixed, predefined patterns or combining strategies (sectors)	It has fixed number of array which can be electronically steered in a particular direction.	An infinite number of patterns (scenario-based) that are adjusted in real time.
2.	This kind of antenna will be easier to implement in existing cell structures than the more sophisticated adaptive arrays, which also means low cost.	Easy to move electronically. In this case, the received power is maximized.	Complex in nature at the time of installment and best performance in the three types of smart antennas.
3.	The signal strength can degrade rapidly during the beam switching.	It does not null the interference.	Excellent performance in interference.

phased array, digital beam forming, adaptive antenna systems, and others. Smart antenna systems are customarily categorized, however, as switched beam, dynamically phased array and adaptive array systems [5].Switched lobe creates a group of overlapping beams that together result in omni directional coverage. The overlapping beam patterns pointing in slightly different directions. The SBSA creates a number of two-way spatial channels on a single conventional channel in frequency, time, or code. Each of these spatial channels has the interference rejection capabilities of the array, depending

on side lobe level [70].As the mobile moves, beam-switching algorithms for each call determine when a particular beam should be selected to maintain the highest quality signal and
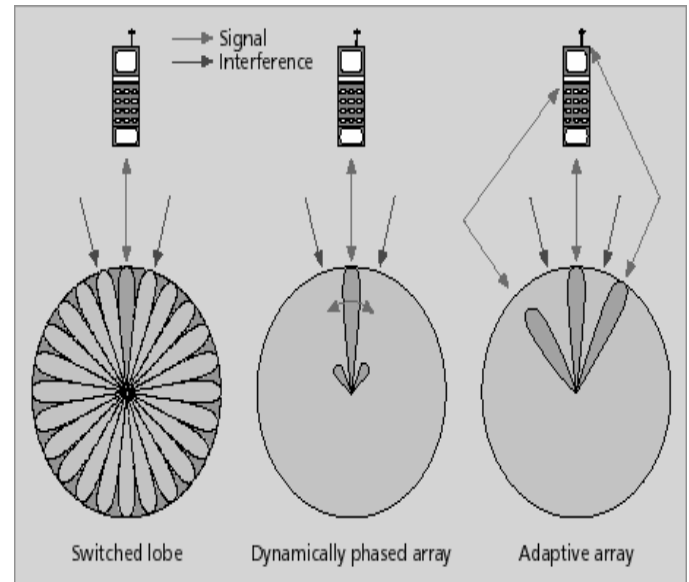
Fig.3. Comparison between three basic types of smart antenna.

the system continuously updates beam selection, ensuring that user gets optimal quality for their call. The system scans the outputs of each beam and selects the beam with the largest output power as well as suppresses interference arriving from directions away from the active beam's center.[70]

The dynamically phased array smart antenna is an antenna which controls its own pattern by means of feed-back or feed-forward control, and it performs gain enhancement for desired signals whereas suppression for interfering signals The phased array antenna consists of   multiple stationary antenna elements, which are fed coherently and use variable phase or time delay control at each element to scan a beam to given angle in space. Array can be used in place of fix aperture antennas(reflectors , lenses), because the multiplicity of elements allows more precise control of radiation pattern, thus resulting in lower side band and careful pattern shaping .

The adaptive array system required sophisticated signal processing algorithm to distinguish between desired signal , multipath signal and interference signal. It combine adaptive digital signal processing to the spatial signal processing to achieve greater performance.

III.   BEAMFORMING SCHEMES OF SMART ANTENNA ARRAY

The Beamforming scheme is important factor to convert antenna array into smart antenna. These schemes tilt the radiation pattern into desired direction depending upon conditions. The simplest beamformer has all the weights of equal magnitudes, and is called a conventional Beamformer or a delay-and sum beamformer. This array has unity response in the look direction, which means that the mean output power of the processor, due to a source in the look

direction, is the same as the source power to steer the array in a particular direction, the phases are selected appropriately. This beamformer provides the maximum output SNR for the case that no directional jammer operating at the same frequency exists, but it is not effective in the presence of directional jammers, intentional or unintentional. Generally null steering and optimal Beamformer are the commonly used in Smart antenna array .

A. Null-Steering Beamformer

Null-steering beamforming techniques require not only control of phase (as for conventional beamforming), but also independent control of the amplitude. A null-steering Beam former can cancel a plane wave arriving from a known direction, producing a null in the response pattern in this direction. The process works well for canceling strong interference, and could he repeated for multiple-interference cancellation. But although it is easy to implement for signal interference, it becomes cumbersome as the number of interference grows. Although the beam pattern produce by this Beamformer has nulls in the directions of interference [5], it is not designed to minimize the uncorrelated noise at the array output. This can be achieved by selecting weights that minimize the mean output power, subject to the above constraints. The flexibility of array weighting to being adjusted to specify the array pattern is an important property. This may be exploited to cancel directional sources operating at the same frequency as that of the desired source, provided these are not in the direction of the desired source. In situations where the directions of these interferences are known, cancellation is possible by placing the nulls in the pattern corresponding to these directions and simultaneously steering the main beam in the direction of the desired signal. Beam forming in this way, where nulls are placed in the directions of interferences, is normally known as null beam forming or null steering. The cancellation of one interference by placing a null in the pattern uses one degree of the freedom of the array. Null beam forming uses the directions of sources toward which nulls are placed for estimating the required weighting on each element. There are other schemes that do not require directions of all sources. A constrained Beamforming scheme uses the steering vector associated with the desired signal and then estimates the weights by solving an optimization problem. Knowledge of the steering vector associated with the desired signal is required to protect the signal from being canceled. In situations where the steering vector associated with the signal is not available, a reference signal is used for this purpose [54].

B. Optimal Beamformer

The optimal Beamformer referred also as the optimal combiner or minimum variance distortion less response beam former (MVDR), does not require knowledge of the direction and the power level of interference ,nor the level of the background noise power , to maximize the output SNR. In this case the weights are computed assuming all source as interference and processor is referred to as a noise along matrix inverse(NAMI) or maximum likelihood (ML) filter ,as it finds the ML estimate of the power of the signal source with the above assumption. Minimizing the total output noise, while keeping the output signal constant, is the same as

maximizing the output SNR. This method requires the number of interferers to be less than or equal to L -2, as an array with L elements has L- 1 degrees of freedom, and one has been utilized by the constraint in the look direction. This may not he true in a mobile-communications environment with multi-path arrivals, and the array Beamformer may not be able to achieve the maximization of the output SNR by suppressing every interference. However, the Beamformer does not have to fully suppress interference, since an increase of a few
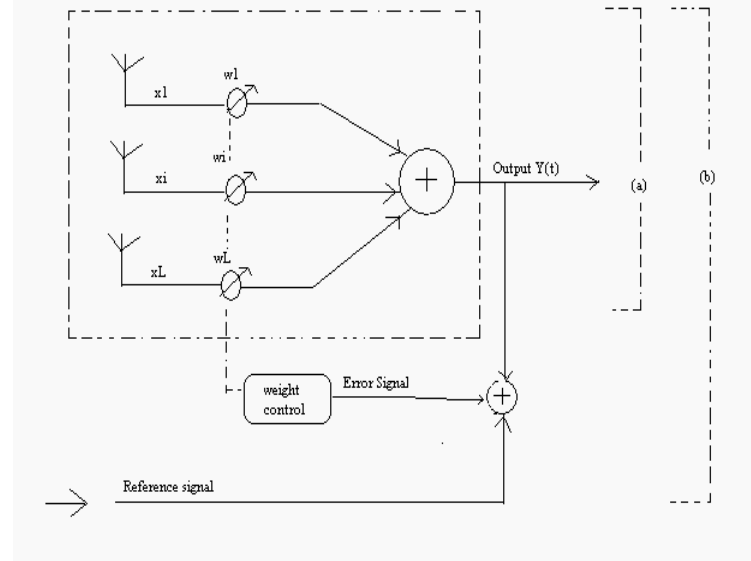
Fig 4 The structure of a narrow band beam-former[10] (a)without reference signal.and (b) using a reference signal.

decibels in the output SNR can make a large increase in the channel capacity. In the optimization using reference signal method, the processor requires a reference signal instead of the desired signal direction (Fig.4). The array output is subtracted from an available reference signal to generate an error signal, which is used to control the weights. Weights are adjusted such that the mean squared error (MSE) between the array output and the reference signal is minimized. Arrays which use zero reference signals are referred to as power-inversion adaptive arrays. The MSE minimization sachem is a closed-loop method, compared to the open –loop scheme of MVDR (the ML filter), and the increased SNR is achieved at the cost of some signal distortion, caused by the filter.

IV.   GENERALLY USED SMART ANTENNA ALGORITHMS

At present, there are many sorts of algorithms that can be applied to the smart antenna systems. People also put forward many modified algorithms on the basis of the basic algorithms to adapt to different performance demands. Generally, there are two categories: blind algorithm and non blind algorithm. The algorithm that needs the reference signal to adjust the weights gradually is referred to as the blind algorithm. Besides, when the directions of the signals are known, we can determine the channel response firstly, and then determine the weights according to certain principle.

This kind of algorithms includes LMS, RLS, SMI, LCMV and so on. Inversely, the blind algorithm doesn't need the reference signal. The receiver can estimate the transmitted signal and treat it as the reference signal to make signal processing. This kind of algorithm makes use of the inherent characteristics of the modulating signal or the characteristics that is independent of the carried information. This kind of algorithms includes CMA, subspace algorithm, MUSIC algorithm, and so on. Moreover, the two kind of algorithm can also be combined, namely, using the non blind algorithm to determine the initial value and then using the blind algorithm to track and adjust, such as SMI+CMA[l]. This method is suitable to the communication system that transmits the pilot symbols.

A. LMS Algorithm

The LMS algorithm is based on the principle of the steepest descend and is applied to the MSE performance measurement. The LMS algorithm intrudes three categories [52] unconstrained LMS algorithm, normalized LMS algorithm and constrained LMS algorithm. When the weights are subjected to constraints at each iteration, the algorithm is referred to as the constrained LMS algorithm. Otherwise, it is referred to as an unconstrained LMS algorithm. The unconstrained LMS algorithm is mostly applicable when weights are updated using a reference signal and no knowledge of the direction of the signal is utilized. Though the structure of the normal LMS algorithms are very simple, it doesn't perform well due to its slow convergence rate in situation of fast-changing signal characteristics and the high sensitivity to the eigen value distribution of the covariance matrix of the array signals, which limits its application in CDMA system. The normalized LMS algorithm is a variation of the constant-step-size LMS algorithm, and uses a data-dependent step size at each iteration .

$$\mu(n) = \frac{\mu}{\overline{X}^H(n)\,\overline{X}(n)}$$

(1)

The algorithm normally has better convergence performance and less signal sensitivity compared to the normal LMS algorithm. When applied to the multi-antenna CDMA mobile systems, using an optimal step-sequence in the update, the algorithm can achieve a fast convergence and a near-optimum steady-state performance at the expense of low increase in the complexity than the normal LMS algorithm[53]. Moreover, a modified and normalized. LMS (MN-LMS) algorithm is presented in [43]. The adaptive filter using this algorithm can track the individual total input phase at each element and the channel estimation and phase calibration are not required for the inverse link improvement.

B. RLS Algorithm

The RLS algorithm is based on the LS rule to make the error square-sum of the array output in each snapshot least . This algorithm take advantage of all the array data information that obtained after the initiation of the algorithm and using the iteration method to realize the inverse operation of the matrix, so the convergence rate is rapid and can realize the tradeoff between the rate of the convergence and the computing

complexity. This algorithm is not sensitive to the eigen value distribution, but compared to the normal LMS algorithm, its computational complexity is high[54].The common solution of the algorithm is

$$\vec{W}(K+1) = \vec{W}(K) - \frac{\overline{P}(K)e_k^{\prime}\,\overline{x}(k+1)}{\mu + \overline{x}(k+1)\overline{x^H}(k+1)\overline{p(k)}}$$

(2)

Where the inverse matrix is updated as

$$\overline{P}(K+1) = \frac{1}{\mu}\left\{ \overline{P}(K) - \frac{\overline{P}(K)\overline{x}(k+1)\overline{x^H}(k+1)\overline{p(k)}}{\mu + \overline{x}(k+1)\overline{x^H}(k+1)\overline{p(k)}} \right\}$$

(3)

Where $\overline{P}(K) = \overline{R^{-1}}(K)$

C. Sample Matrix Inversion (SMI) Algorithm

The SMI algorithm estimates the weights directly by estimating the covariance matrix R from K independent samples of data by time- averaging. Thus the problem that the rate of the convergence depends on the eigen value distribution can be avoided. The optimum solution obtained from the SMI algorithm is[55].

$$\overline{W} = \overline{R}^{-1} V$$

(4)

Where $\overline{R} = \frac{1}{K}\sum \overline{X_i X_i}^H$

$\overline{X}_i$ is a complex sample vector of receiver outputs of length N, N is the number of elements of the array antenna, K is the number of sample vectors used. V is a steering vector of length N which is equal to the un adapted array weights. Forming a sample covariance matrix and solving for the weights provides a very fast rate of convergence. The rate of convergence is dependent only on the number of elements and is independent of the noise and interference environment and the eigen value distribution. Because the complexity of the computing is proportional to N3 so it requires that the algorithm has a strong processing ability when the array is large. To a certain given value of K, the quality of the estimation obtained from the time average is dependent on the input signal-noise ratio (SNR). When the SNR decreases, in order to eliminate noise and interference, a large amount of samples are needed to obtain the estimation more precisely . Ronald L. etc had ever put forward the M-SMI algorithm[66], namely the modified SMI, in which the diagonal loading technique is used, where, the diagonal of the covariance matrix is augmented with a positive or negative constant prior to inversion. Compared to the SM1 algorithm, the diagonally loaded sample covariance matrix

$$\overline{\overline{R}} = \overline{R} + F\overline{I}$$

(5)

F can be positive or negative, but for the covariance matrix to be positive definite. The positive loading tends to reduce the null depth on weak interfering signal, while it

decreases the convergence time. Conversely, negative loading tends to increase the null depth on weak interfering signals while increasing convergence time. The SMI algorithm can get the maximum signal-to-interference-plus-noise (SINR). However, in some applications, such as digital communications or satellite television communications, other measures of performance such as SIR may be equally important, the M-SMI can be applied in this situation.

D. LCMV Algorithm

The algorithms mentioned above all need the reference signal, and the reference signal must have Large correlation with the desired signal. But in actual environment, this is difficult to obtain. So we can make use of the technology of orientation of the reference signal source. In the environment that the signals are dense, we can orient the desired signal and the interference signal sources, and then combine this with the technology of nulling adaptively, thus we can obtain reference nulling with high resolution. It is assumed that there are p desired signals and q interference signals incident on the antenna. The directions of the incident signals are (θ_i ….., θ_p) and (θ_{p+1} …….., θ_{p+q}) respectively in which p + q < M . The constrained condition of the LCMV algorithm[57] is:

$$\begin{cases} \min_{\vec{W}} P(\vec{W}) = \min_{\vec{W}}(\vec{W}^H \vec{R}_x \vec{W}) \\ \vec{A}^H \vec{W} = \vec{e} \end{cases}$$

(6)

Where

$$A = \left[\vec{a}(\theta_1), \cdots \vec{a}(\theta_p), \vec{a}(\theta_{p+1}), \cdots \vec{a}(\theta_{p+q})\right],$$

$$\vec{e} = \left[\underbrace{1, \cdots, 1}_{p}, \underbrace{0, \cdots 0}_{q}\right]^H$$

This algorithm can ensure that the antenna has the gain of 1 in the directions of the desired signals, while the responses in the directions of the interference signals are zero, thus there are deep nulls in the directions of the interference signals, which can be seen from the directional pattern of the antenna, Through these constrained conditions, the interference signals can be suppressed and the output power of the array can be minimized to suppress other signals and noises which are not located in the main lobe of the antenna. The weight vector of the LCMV algorithm is:

$$W = R_{\bar{x}}^{-1} A (A^H R_{\bar{x}}^{-1} A)^{-1} \vec{e}$$

(7)

From above equation, we can see that in DS-CDMA systems, the above two algorithms, namely SMI and LCMV algorithms, can be used by the adaptive antenna array for propagation delay estimation. The large sample maximum likelihood (LSML) is applied to the beam forming output data for estimating to the propagation delay of a desired user in multi-user sceneries. The adaptive antenna array can help the LSML estimator to obtain improved performances as compared to a single antenna based LSML estimator.

E. CMA Algorithm

In order to adaptively control directions of nulls, some information concerning incident waves such as directions and intensity of incident waves is required. It is , however, very difficult to know the information in some environment. In addition, the directions and intensity may vary with the variation of the environment. Thus the algorithm for controlling the nulls is important especially in the case of above environment. The CMA algorithm can solve the problem [58]. It is a typical blind algorithm and only requires that the amplitude of the transmitted signal is constant, such as FM, PSK, FSK etc. CMA is based on the fact that the amplitude of the combined signal fluctuates because of the interference. Thus, in CMA. the amplitude of the combined signal is always observed, and the weights coefficients are adjusted so as to minimize the variation of the amplitude of the signal. When the output amplitude becomes constant, nulls can be formed in the direction of the interference signals on the directional pattern. Moreover, Satoshi Denno etc have put forward the Modified CMA algorithm in [59].The use of adaptive array to reject wideband interferences and track wideband signals has been proven to be more efficient if frequency compensation is used. Among the frequency compensation algorithms, the interpolating techniques have been applied to the CMA. ICMA permits to improve system performances by readjusting the main lobe's direction toward the signal's DOA and increasing the interference null depth [60].

V. FURTHER REMARKES

In this paper, we have discussed various Smart antenna array architectures, Beamforming techniques and algorithms. The design and architecture of smart antenna is case sensitive and changed according to the demand of applications. The adaptive array provide excellent result in the presence of interference, but its design is more complex and costly as compared of other two. In Beamforming null steering Beamforming perform well in case of strong interferences, but in need prior information of that. The blind algorithm doesn't need the reference signal so we can apply them according the communication system demands.

REFERENCES

[1] Fang-Biau Ueng, Jun-Da Chen and Sheng-Han Cheng "Smart Antennas for Multi-user DS/CDMA communications in Multipath Fading Channels" IEEE Eighth international symposium on spread spectrum ISSSTA2004, Sydney, Australia, 30 Aug. - 2 Sep. 2004

[2] Alexiou, A. , Haardt, M. "Smart antenna technologies for future wireless systems: trends and challenges " IEEE Communications Magazine,Volume 42, Issue 9, Page(s):90 - 97 Sept. 2004

[3] A. Rawat "Smart antenna terminal development" National conf. of IETE Chandigadh, India April 2005

[4] A. Rawat "Design of smart antenna system for military application using mat lab" National conf. of Institution of Engineers in Jaipur , India Aug 2006.

[5] Chryssomallis, M." smarty antennas" Antennas and Propagation Magazine, IEEE Volume 42, Issue 3, Page(s):129 – 136 June 2000

[6] A. Paulraj, R. Nabar, and D. Gore, "Introduction to Space-Time Wireless Communications", Cambridge Univ. Press, 2003.

[7] L. C. Codara, "Application of Antcnna Arrays to Mobile Communications, Part 11: Beam-Forming and Direclion-of-Arrival Considerations," Proceedings of the IEEE, 85, pp.1195-1245, 8, August 1997

[8] Lal C. Godra, Application of Antenna Array to Mobile Communications, Part U : Beam-Forming and Direction-of-Arrival

Considerations". Proceedings of the IEEE, Vol. 85, No. 8, Page(s): 1213-1218, 1997.

[9] Jian-Wu Zhang "The Adaptive Algorithms of the Smart Antenna System in Future Mobile Telecommunication Systems" IEEE International Workshop on Antenna Technology pp347-350, 2005

[10] Blair D. Carlson, "Covariance Matrix Estimation Erron and Diagonat Loading in Adaptive Arrays". IEEE Transactions on Aerospace and Electronic System. Vol. 24, No. 4, Page(s): 397-401, July 1988.

[11] Werner, S.; Apolinario, J.A.., Jr.; Lakkso, T.I. "Multiple-antenna CDMA Mobile Reception Using Constrained Normalized Adaptive Algorithms", Telecommunications Symposium, 199s. ITS '98 Proceedings. SBT/IEEE International, Vol: 1, Page(s): 353-358 , 1998 .

[12] Fujimoto, M.; Nishikawa, K.; Sato, K., "A Study of Adaptive Array Antenna System for Land Mobile Communications", Intelligent Vehicles'95 Symposium, Proceedings of the IEEE, Page(s): 36-41, 25-26 Sept, 1995.

[13] Demo, S.; Ohira, T., "M-CMA for Digital Signal Processing Adaptive Antennas with Microwave Beamforming",Proceedings of IEEE, Vol. 5, Page(s): 179-187 ,2000 .

[14] Hefnawi, M.; Delisle, G.Y. "Adaptive arrays for wideband interference suppression in wireless communications", Antennas and Propagation Society, 1999. IEEE International Symposium 1999, vok3, Page(s): 1588 - 1591, 1999.

[15] Weijun Yao, and Yuanxun Ethan Wang, "Beamforming for Phased Arrays on Vibrating Apertures", IEEE Trans. Antennas Propag., vol. 54,no.10, Oct. 2006

[16] A. H. El Zooghby, C. G. Christodoulou, and M. Georgiopoulos "Neural Network-Based Adaptive Beamforming for One- and Two-Dimensional Antenna Arrays" IEEE Trans. Antennas Propag., vol. 46, no. 12 pp1891 -1893, Dec. 1998.

[17] Hugh L. Southall,Jeffrey A. Simmers, and Teresa H. O'Donnell "Direction Finding in Phased Arrays with a Neural Network Beamformer" IEEE Trans. Antennas Propag., vol. 43,no. 12 pp 1369-1374 , Dec1995.

[18] Robert J. Mailloux "Phased array antenna handbook" Artech House,2006 .

[19] Eric Charpentier, and Jean-Jacques Laurin, "An Implementation of a Direction-Finding Antenna for Mobile Communications Using a Neural Network" IEEE Trans. Antennas Propag., vol. 47, NO. 7pp 1152 -1158 , JULY 1999.

[20] B. K. Yeo and Y. Lu, "Array failure correction with a genetic algorithm,"IEEE Trans. Antennas Propag., vol. 47, no. 5, pp. 823–828,1999.

[21] M. Salazar-Palma, T. K. Sarkar, L.-E. G. Castillo, T. Roy, and A. Djordjevic , Iterative and Self- Adaptive Finite-Elements in Electromagnetic Modeling. Norwood, MA: Artech House, 1998.

[22] Amalendu Patnaik, B. Choudhury, P. Pradhan, R. K. Mishra, and Christos Christodoulou "An ANN Application for Fault Finding in Antenna Arrays " IEEE Trans. Antennas Propag., vol. 55, no.3pp 775-777, Mar. 2007.

[23] R. F. Harrington, "Field Computation by Moment Methods". New York:IEEE Press, 1993.

[24] L. C. Codara, "Application of Antcnna Arrays to Mobile Communications, Part 11: Beam-Forming and Direclion-of-Arrival Considerations," Proceedings of the IEEE, 85, 8, pp. 1195-1245, August 1997

[25] M. Nagatsuka, N. Ishii, R. Kohno and H. Imai, "Adaptive Array Antcnna Based on Spatial Spcctral Estimation Using Maximum Enlrapy Method," IEICE Trnnsactioris on Corn,rru,ricutiorrs,E77-B, 5, pp. 624-633, 1994.

[26] R. Kohno, C. Yim and H. Imai "Array Antenna Beamforming Based on Estimation on Arrival Angles Using DFT on Spatial Domain," Proceedings of PIMRC 1991, London, UK, ,pp. 38-43 September 1991.

[27] Jumarie, G. "Nonlinear filtering: A weighted mean squares approach and a Bayesian one via the maximum entropy principle."*Signal Processing*, **21** (1990), 323—338, 1990.

[28] Sun, Q., Alouani, A. T., Rice, T. R., and Gray, J. E. " Linear system state estimation: A neurocomputing approach." In *Proceedings of the American Control Conference*, 550—554, 1992.

[29] Cohen, S. A. "Adaptive variable update rate algorithm for tracking targets with a phased array radar'. *IEE Proceedings*, pt. F, 133, 277—280, 1986 .

[30] J.C. Liberti, T.S. Rappaport, "Smart Antennas forWireless Communications: IS-95 and Third-Generation CDMA Applications",Prentice Hall, NJ, 1999.

[31] LAL C. GODAR4, Application of Antenna Array to Mobile Communications, Part U : Beam-Forming and Direction-of-Arrival Considerations". Proceedings of the IEEE, Vol. 85, No. 8, Page(s): 1213-1218, 1997.

[32] Sandchoon Kim; Miller, S.L. "An Adaptive Antenna array Based Propagation Delay Estimation for DS-CDMA Communication Systems", Military Communications Conference, 1998. Milcom 98, Proceedings of the IEEE Vol: 1, Page(s):333-337, 1998 .

[33] Sandchoon Kim; Miller, S.L. "An Adaptive Antenna array Based Propagation Delay Estimation for DS-CDMA Communication Systems", Military Communications Conference, 1998. Milcom 98, Proceedings of the IEEE Vol: 1, Page(s):333-337, 1998 .

[34] BLAIR D. CARLSON, "Covariance Matrix Estimation Erron and Diagonat Loading in Adaptive Arrays". IEEE Transactions on Aerospace and Electronic System. Vol. 24, No. 4, Page(s): 397-401, July 1988.

[35] Ronald L. Dilsavor, Randolph L. Moses, "Analysis of Modified SMI method for adaptive Array Weight Control', IEEE Transactions on Signal Processing, Vol. 41, No. 2, Page(s): 721-726,1993,.

[36] Werner, S.; Apolinario, J.A.., Jr.; Lakkso, T.I. "Multiple-antenna CDMA Mobile Reception Using Constrained Normalized Adaptive Algorithms", Telecommunications Symposium, 199s. ITS '98 Proceedings. SBT/IEEE International, Vol: 1, Page(s): 353-358 , 1998

[37] Fujimoto, M.; Nishikawa, K.; Sato, K., "A Study of Adaptive Array Antenna System for Land Mobile Communications", Intelligent Vehicles'95 Symposium, Proceedings of the IEEE, Page(s): 36-41, 25-26 Sept, 1995.

[38] Demo, S.; Ohira Demo, S.; Ohira, T., "M-CMA for Digital Signal Processing Adaptive Antennas with Microwave Beamforming",Proceedings of IEEE, Vol. 5, Page(s): 179-187 ,2000 .

[39] Hefnawi, M.; Delisle, G.Y. "Adaptive arrays for wideband interference suppression in wireless communications", Antennas and Propagation Society, 1999. IEEE International Symposium 1999, vok3, Page(s): 1588 - 1591, 1999.

Comments on Five Smart Card Based Password Authentication Protocols

Yalin Chen
Institute of Information Systems and
Applications, NTHU, Tawain
d949702@oz.nthu.edu.tw

Jue-Sam Chou[*]
Dept. of Information Management
Nanhua University, Taiwan
jschou@mail.nhu.edu.tw
[*]: corresponding author

Chun-Hui Huang
Dept. of Information Management
Nanhua University, Taiwan
g6451519@mail.nhu.edu.tw

Abstract¡ In this paper, we use the ten security requirements proposed by Liao et al. for a smart card based authentication protocol to examine five recent work in this area. After analyses, we found that the protocols of Juang et al.¡s , Hsiang et al.¡s, Kim et al.¡s, and Li et al.¡s all suffer from offline password guessing attack if the smart card is lost, and the protocol of Xu et al.¡s is subjected to an insider impersonation attack.

Keywords- password authentication protocol; insider attack; smart card loss problem; password guessing attack

I. INTRODUCTION

Password authentication protocols have been widely adopted for a user to access a remote server over an insecure network. In recent, many smart card password authentication protocols [1-20] are proposed, which emphasizes two-factor authentication mechanism to enhance the user end¡s security. One factor is the user-rememberable password while the other factor is the user-possessing smart card which is a tamper-resistant device with storage and computational power. Moreover, recent studies investigated a weakness of a traditional password authentication protocol. That is, in the traditional one the server usually maintains a password or verification table to store user authentication data. However, this approach will make the system easily subjected to impersonation or stolen-verifier attack if the table is compromised.

In 2006, Liao et al. [2] identified ten security requirements to evaluate a smart card based password authentication protocol. We show them as follows.

R1. It needs no password or verification table in the server.

R2. The client can choose and change his password freely.

R3. The client needs not to reveal their password to the server even in the registration phase.

R4. The password should not be transmitted in plaintext over the network.

R5. It can resist insider (a legal user) attack.

R6. It can resist replay attack, password guessing attack, modification-verification-table attack, and stolen-verifier attack.

R7. The length of a password should be appropriate for memorization.

R8. It should be efficient and practical.

R9. It should achieve mutual authentication.

R10. It should resist offline password guessing attack even if the smart card is lost.

In their article, they also proposed a protocol to satisfy these ten security requirements. But Xiang et al. [9] demonstrated that their protocol suffers from both the replay attack and the password guessing attack. Other than theirs, many efforts trying to propose a secure protocol were made recently. For example in 2008, Juang et al. [7] proposed an efficient password authenticated key agreement using bilinear pairings. In 2009, Hsiang et al. [14], Kim et al. [16], and Xu et al. [18] each also proposed a protocol of this kind, respectively. In this year 2010, Li et al.[20] also proposed a protocol in this area. Although they claimed their protocols are secure. However, in this paper, we will show some weaknesses in [18], [7], [14], [16], [20], correspondingly.

The remainder of this paper is organized as follows: In Section II, we review and attack on the scheme of Juang et al.¡s [7]. Then we review and attack on the protocols of Hsiang et al. ¡s [14], Kim et al. [16], Xu et al. ¡s [18], and Li et al. ¡s [20] in Section III through VI, respectively. Finally, a conclusion is given in Section VIII.

II. REVIEW AND ATTACK ON JUANG ET AL.'S SCHEME

In their scheme [7], if an attacker gets C¡s smart card, he can successfully launch an offline password guessing attack. Hence, the scheme cannot satisfy requirement R10. In the following, we first review Juang *et al.*¡s protocol and then show the attack on the protocol.

A. Review

Their protocol consists of four phases: the setup phase, the registration phase, the login and authentication phase, and the password changing phase.

In the setup phase, server S chooses two secrets s, x and publishes $P_s = sP$, where P is a generator of an additive cyclic

group G_1 with a prime order q. S also publish a secure hash function H(¡).

In the registration phase, user i register his ID_i and H(PW_i, b) to server S. S issues a smart card which contains b_i (b_i = E_x[H(PW_i, b), ID_i, H(H(PW_i, b), ID_i)], E_x[M] which is a ciphertext of M encrypted by S¡s secret key x), and b (a random number chosen by i).

When i wants to login into S, i starts the login and authentication phase, and sends {aP, α} to S, where a is a random number chosen by i, α = $E_{Ka}[b_i]$, Ka = H(aP, P_s, Q, e(P_s, aQ)), e: G_1¡ $G_1 \rightarrow G_2$ is a bilinear mapping, Q = h(ID_s), h(¡) is a map-to-point hash function, h:$\{0,1\}^* \rightarrow G_1$, and ID_s is S¡s identification. Subsequently, S chooses a random number r, computes the session key sk = H(H(aP, P_s, Q, e(aP, sQ)), r, ID_i, ID_s) = H(Ka, r, ID_i, ID_s) since e(P_s, aQ) = e(aP, sQ) , and sends {$Auth_s$, r} to user i, where $Auth_s$ = H(Ka, H(PW_i, b), r, sk), and H(PW_i, b) is obtained from decrypting α and b_i. Then, i computes the session key sk. To authenticate S, user i verifies $Auth_s$ to see if it is equal to H(Ka, H(PW_i, b), r, sk). If it is, i computes and sends {$Auth_i$} to S, where $Auth_i$ = H(Ka, H(PW_i, b), $r+1$, sk) and H(PW_i, b) is the hash result of b stored in the smart card with PW_i inputted by i. Finally, to authenticating i, S checks to see if $Auth_i$ is equal to H(Ka, H(PW_i, b), $r+1$, sk).

B. Attack

In the protocol, supposed that user C lost his smart card and the card is got by an insider E, E can impersonate C to login into S without any detection. We show the attack in the following.

E first reads out b and b_c (which equals E_x[H(PW_c, b), ID_c, H(H(PW_c, b), ID_c)]) stored in C¡s smart card but he doesn¡t have the knowledge of PW_c.

In the login and authentication phase, E chooses a random number c, computes cP, Kc = H(cP, P_s, Q, e(P_s, cQ)), α = $E_{Kc}[b_c]$, and sends {cP, α} to S. After receiving the message, S chooses a random number r, computes session key sk = H(Kc, r, ID_c, ID_s), $Auth_s$ = H(Kc, H(PW_c, b), r, sk), and sends {$Auth_s$, r} to C. E intercepts the message and launches an off-line password guessing attack as follows.

E chooses a candidate password PW' from a dictionary, computes Kc = H(cP, P_s, Q, e(P_s, cQ)), sk = H(Kc, r, ID_c, ID_s), H(Kc, H(PW', b), r, sk) and checks to see if it is equal to the received $Auth_s$. If it is, the attacker successfully gets C¡s password PW_c which is equal to PW'. Subsequently, E can masquerade as C by using PW' and C¡s smart card to log into S. That is, Juang et al.¡s cannot satisfy the security requirement R10: It should resist password guessing attack even if the smart card is lost.

III. Review and attack on the protocol of Hsiang et al.'s scheme

In this section, we first review Hsiang et al.¡s protocol [14] and then demonstrate a smart card lost and offline password guessing attack on the protocol.

A. Review

In the protocol, when user C wants to change his password, he inserts his card and types his ID and PW. The smart card computes P^* = $R \oplus$ H($b \oplus PW$), and V^* = H($P^* \oplus$ H(PW)), and compares V^* with V, where PW is C¡s old password, and R, b, and V are stored in C¡s smart card. If they are equal, the card verifies user C and accepts his password change request. The card subsequently ask C a new password PW^* and then computes R_{new} = $P^* \oplus$ H($b \oplus PW^*$) and V_{new} = H($P^* \oplus$ H(PW^*)). Finally, the card replaces V with V_{new}.

B. Attack

Assume that an attacker E who gets C¡s smart card, reads the values of R, b, and V, and then launches an offline password guessing attack as follows. E chooses a candidate password PW' from a dictionary, computes P' = $R \oplus$ H($b \oplus PW'$) and V' = H($P' \oplus$ H(PW')), and checks to see if V' and V are equal. If they are, PW' is the correct password.

IV. Review and attack on the protocol of Kim et al.'s scheme

In this section, we first review Kim et al.¡s protocol [16] and then demonstrate a smart card lost and offline password guessing attack on the protocol.

A. Review

In their protocol, when user C wants to change his password, he inserts his card and types his ID and PW. The smart card computes K^*_1 = $R \oplus$ H(PW) and compares K^*_1 with K_1 to see if they are equal, where R ($=K_1 \oplus$ H(PW_c)) and K_1 ($=$H($ID \oplus x$) $\oplus N$) are stored in C¡s smart card, PW_c is chosen by the user when he registers himself to the remote server S, and N is a random number. If they are equal, the card verifies user C and accepts his password change request. C subsequently asks C a new password PW^*, and then computes R^* = $K^*_1 \oplus$ H(PW^*) and K^*_2 =$K_2 \oplus$ H($PW \oplus$ H(PW)) $\oplus$ H($PW^* \oplus$ H(PW^*)), where K_2 = H($ID \oplus x \oplus N$) $\oplus$ H($PW_c \oplus$ H(PW_c)) is also stored in C¡s smart card. Finally, the smart card will replace R and K_2 with R^* and K^*_2, respectively.

B. Attack

An attacker E who gets C¡s smart card, reads the values of R, K_1, and K_2, and then launches an offline password guessing attack as follows. E chooses a candidate password PW' from a dictionary, computes K'_1 = $R \oplus$ H(PW'), and checks to see if K'_1 and K_1 are equal. If they are, PW' is the correct password.

V. Review and attack on the protocol of Xu et al.'s scheme

Xu et al.¡s protocol [18] can not satisfy security requirements R3 (The client needs not to reveal their password to the server) and R5 (It can resist insider attack). We show the scheme and its violations as follows.

A. Review

Xu *et al.*¡s protocol [18] consists of three phases: the registration phase, the login phase, and the authentication phase.

In the registration phase, user C submits his ID_c and PW_c to the server S. S issues C a smart card which stores C¡s identity ID_c, and $B = H(ID_c)^x + H(PW_c)$, where x is S¡s secret key and PW_c is C¡s password.

In the login phase, user C inputs ID_c and PW_c to his smart card. The card obtains timestamp T, chooses a random number v, computes $B_c = (B_¡ H(PW_c))^v = H(ID_c)^{xv}$, $W = H(ID_c)^v$, and $C_1 = H(T, B_c, W, ID_c)$, and sends $\{ID_c, C_1, W, T\}$ to S.

In the authentication phase, after receiving $\{ID_c, C_1, W, T\}$ at time T^*, S computes $B_s = W^x$, and checks to see if ID_c is valid, $T^* - T < \Delta T$, and C_1 is equal to $H(T, B_s, W, ID_c)$. If they are, S selects a random number m, gets timestamp T_s, computes $M = H(ID_c)^m$, $C_s = H(M, B_s, T_s, ID_c)$, and sends $\{ID_c, C_s, M, T_s\}$ to C. After receiving the message, C verifies ID_c and T_s, computes $H(M, B_c, T_s, ID_c)$, and compares it with the received C_s. If they are equal, S is authentic. Then, C and S can compute the common session key as $sk = H(ID_c, M, W, M^v)$ and $sk = H(ID_c, M, W, W^m)$, respectively.

B. Weaknesses

First, the scheme obviously violates security requirement R3 since the client transmits clear password in the registration phase.

Second, we show an impersonation attack on the scheme below. Assume that a malicious insider U wants to masquerade as C to access S¡s resources. He reads B from his smart card, obtains system¡s timestamp T_u, chooses a random number r, computes $B_u = (B_¡ H(PW_u))^r = H(ID_u)^{xr}$, $W = H(ID_c)^r$, $C_1 = H(T_u, B_u, W, ID_c)$, and sends $\{ID_c, C_1, W, T_u\}$ to S.

After receiving the message, S validates ID_c and T_u, computes $B_s = W^x = H(ID_c)^{rx}$, and checks to see if the received C_1 is equal to the computed $H(T_u, B_s, W, ID_c)$. In this case, we can see that C_1 is obviously equal to $H(T_u, B_s, W, ID_c)$. Hence, U (who masquerades as C) is authentic. Finally, S obtains timestamp T_s and sends $\{ID_c, C_s, M, T_s\}$ to U, where $M = H(ID_c)^m$ and m is a random number chosen by S. U also can compute the session key as $sk = H(ID_c, M, W, M^r)$ shared with S. Therefore, user U¡s insider impersonation attack succeeds.

VI. REVIEW AND ATTACK ON THE PROTOCOL OF LI ET AL.'S SCHEME

In this section, we first review the registration phase, login phase and authentication phase of the protocol in Li *et al.*¡s [20], and then present our attack on the protocol.

A. Review

In the registration phase, user C submits his ID_c, PW_c, and his personal biometric B_c to the server S. S issues a smart card for C, which stores the values of ID_c, $f_c = H(B_c)$, and $e_c = H(ID_c, x) \oplus H(PW_c, f_c)$, where x is S¡s secret key.

In the login phase, user C keys ID_c and PW_c to his smart card and inputs his personal biometric B_c on the specific device to check if $H(B_c)$ is equal to f_c stored in the smart card. If it is, the card selects a random number R_c, computes $M_1 = e_c \oplus H(PW_c, f_c) = H(ID_c, x)$, $M_2 = M_1 \oplus R_c$, and sends $\{ID_c, M_2\}$ to S.

In the authentication phase, after receiving $\{ID_c, M_2\}$, S checks to see if ID_c is valid. If it is, S chooses a random number R_S, computes $M_3 = H(ID_c, x)$, $M_4 = M_2 \oplus M_3 = R_c$, $M_5 = M_3 \oplus R_S$, $M_6 = H(M_2, M_4)$, and sends $\{M_5, M_6\}$ to C. After receiving S¡s message, C verifies whether M_6 is equal to $H(M_2, R_c)$. If it is, S is authentic. C then computes $M_7 = M_5 \oplus M_1 = M_3 \oplus R_S \oplus M_1 = H(ID_c, x) \oplus R_S \oplus H(ID_c, x) = R_S$, $M_8 = H(M_5, M_7)$, and sends $\{M_8\}$ to S. After receiving C¡s message, S verifies whether M_8 is equal to $H(M_5, R_s)$. If it is, C is authentic. S then accepts C¡s login request.

B. Attack

Assume that an attacker E gets C¡s smart card and reads the values of ID_c, f_c and e_c. He can launch an offline password guessing attack by sending only one login request to the server. We show the attack as follows.

E chooses a random number M_e and sends $\{ID_c, M_e\}$ to S. After receiving the message, S checks to see if ID_c is valid. If it is, S chooses a random number R_S, computes $M_3 = H(ID_c, x)$, $M_4 = M_e \oplus M_3$, $M_5 = M_3 \oplus R_S$, $M_6 = H(M_e, M_4)$, and sends $\{M_5, M_6\}$ to E. After receiving S¡s message, E terminates the communication, chooses a candidate password PW' from a dictionary, computes $M' = H(M_e, M_e \oplus e_c \oplus H(PW', f_c))$, and compares to see if M' is equal to M_6. If they are, PW' is the correct password, since $M_e \oplus e_c \oplus H(PW', f_c) = M_e \oplus H(ID_c, x) \oplus H(PW_c, f_c) \oplus H(PW', f_c)$. If $PW' = PW_c$, then the equation equals to $M_e \oplus H(ID_c, x)$ which equals to $M_e \oplus M_3 = M_4$. That is, $M' = H(M_e, M_4) = M_6$.

VII. CONCLUSION

Smart-card based password authentication protocols provide two-factor authentication mechanism to improve the user end¡s security than the traditional ones. Liao et al. proposed ten security requirements to evaluate this kind of protocols. According these ten requirements, we investigate recent five schemes. Juang et al.¡s scheme suffers smart card lost and impersonation attack. Kim et al.¡s, Hsiang et al.¡s, and Li et al.¡s schemes are subjected to smart card lost and offline password guessing attack. Finally, Xu et al.¡s scheme has weakness of insider impersonation attack.

REFERENCES

[1] H. Y. Chien, C. H. Chen, ¡A Remote Authentication Preserving User Anonymity,¡ *Proceedings of the 19th International Conference on Advanced Information Networking and Applications* (AINA ¡05), Vol.2, pp. 245-248, March 2005.

[2] I. E. Liao, C. C. Lee, M. S. Hwang, ¡A password authentication scheme over insecure networks¡, *Journal of Computer and System Sciences*, Vol. 72, No. 4, pp. 727-740, June 2006.

[3] T. H. Chen, W. B. Lee, ¡A new method for using hash functions to solve remote user authentication¡, Computers *& Electrical Engineering*, Vol. 34, No. 1, pp. 53-62, January 2008.

[4] C. S. Bindu, P. C. S. Reddy, B. Satyanarayana, ¡Improved remote user authentication scheme preserving user anonymity¡, *International Journal of Computer Science and Network Security*, Vol. 8, No. 3, pp. 62-65, March 2008.

[5] Y. Lee, J. Nam, D. Won, ¡Vulnerabilities in a remote agent authentication scheme using smart cards¡, *LNCS: AMSTA*, Vol. 4953, pp. 850-857, April 2008.

[6] W. S. Juang, S. T. Chen, H. T. Liaw, ¡Robust and efficient password-authenticated key agreement using smart cards¡, *IEEE Transactions on Industrial Electronics*, Vol. 55, No. 6, pp. 2551-2556, June2008.

[7] W. S. Juang, W. K. Nien, ¡Efficient password authenticated key agreement using bilinear pairings¡, *Mathematical and Computer Modelling*, Vol. 47, No. 11-12, pp. 1238-1245, June 2008.

[8] J. Y. Liu, A. M. Zhou, M. X. Gao, ¡A new mutual authentication scheme based on nonce and smart cards¡, *Computer Communications*, Vol. 31, No. 10, pp. 2205-2209, June 2008.

[9] T. Xiang, K. Wong, X. Liao, ¡Cryptanalysis of a password authentication scheme over insecure networks¡, *Computer and System Sciences*, Vol. 74, No. 5, pp. 657-661, August 2008.

[10] G. Yang, D. S. Wong, H. Wang, X. Deng, ¡Two-factor mutual authentication based on smart cards and passwords¡, *Journal of Computer and System Sciences*, Vol. 74, No. 7, pp.1160-1172, November 2008.

[11] T. Goriparthi, M. L. Das, A. Saxena, ¡An improved bilinear pairing based remote user authentication scheme¡, *Computer Standards & Interfaces*, Vol. 31, No. 1, pp. 181-185, January 2009.

[12] H. S. Rhee, J. O. Kwon, D. H. Lee, ¡A remote user authentication scheme without using smart cards¡, *Computer Standards & Interfaces*, Vol. 31, No. 1, pp. 6-13, January 2009.

[13] Y. Y. Wang, J. Y. Liu, F. X. Xiao, J. Dan, ¡A more efficient and secure dynamic ID-based remote user authentication scheme¡, *Computer Communications*, Vol. 32, No. 4, pp. 583-585, March 2009.

[14] H. C. Hsiang, W. K. Shih, ¡Weaknesses and improvements of the Yoon¡ Ryu¡Yoo remote user authentication scheme using smart cards¡, *Computer Communications*, Vol. 32, No. 4, pp. 649-652, March 2009.

[15] D. Z. Sun, J. P. Huai, J. Z. Sun, J. X. Li, ¡Cryptanalysis of a mutual authentication scheme based on nonce and smart cards¡, *Computer Communications*, Vol. 32, No. 6, pp. 1015-1017, April 2009.

[16] S. K. Kim , M. G. Chung, ¡More secure remote user authentication scheme¡, *Computer Communications*, Vol. 32, No. 6, pp. 1018-1021, April 2009.

[17] H. R. Chung, W. C. Ku, M. J. Tsaur, ¡Weaknesses and improvement of Wang et al.'s remote user password authentication scheme for resource-limited environments¡, *Computer Standards & Interfaces*, Vol. 31, No. 4, pp. 863-868, June 2009.

[18] J. Xu, W. T. Zhu, D. G. Feng, ¡An improved smart card based password authentication scheme with provable security¡, *Computer Standards & Interfaces*, Vol. 31, No. 4, pp. 723-728, June 2009.

[19] M. S. Hwang, S. K. Chong, T. Y. Chen, ¡DoS-resistant ID-based password authentication scheme using smart cards¡, *Journal of Systems and Software*, In Press, Available online 12 August 2009.

[20] C. T. Li, M. S. Hwang, ¡An efficient biometrics-based remote user authentication scheme using smart cards¡, *Journal of Network and Computer Applications*, Vol. 33, No. 1, pp. 1-5, January 2010.

AUTHORS PROFILE

Yalin Chen received her bachelor degree in the depart. of computer science and information engineering from Tamkang Univ. in Taipei, Taiwan and her MBA degree in the department of information management from National Sun-Yat-Sen Univ. (NYSU) in Kaohsiung, Taiwan. She is now a Ph.D. candidate of the Institute of Info. Systems and Applications of National Tsing-Hua Univ.(NTHU) in Hsinchu, Taiwan. Her primary research interests are data security and privacy, protocol security, authentication, key agreement, electronic commerce, and wireless communication security.

Jue-Sam Chou received his Ph.D. degree in the department of computer science and information engineering from National Chiao Tung Univ. (NCTU) in Hsinchu, Taiwan,ROC. He is an associate professor and teaches at the department of Info. Management of Nanhua Univ. in Chiayi, Taiwan. His primary research interests are electronic commerce, data security and privacy, protocol security, authentication, key agreement, cryptographic protocols, E-commerce protocols, and so on.

Chun-Hui Huang is now a graduate student at the department of Info. Management of Nanhua Univ. in Chiayi, Taiwan. She is also a teacher at Nantou County Shuang Long Elementary School in Nantou, Taiwan. Her primary interests are data security and privacy, protocol security, authentication, key agreement.

Cryptanalysis on Four Two-Party Authentication Protocols

Yalin Chen
Institute of Information Systems and
Applications, NTHU, Tawain
d949702@oz.nthu.edu.tw

Jue-Sam Chou[*]
Dept. of Information Management
Nanhua University, Taiwan
jschou@mail.nhu.edu.tw
[*]: corresponding author

Chun-Hui Huang
Dept. of Information Management
Nanhua University, Taiwan
g6451519@mail.nhu.edu.tw

Abstract¡ **In this paper, we analyze four authentication protocols of Bindu et al., Goriparthi et al., Wang et al. and Holbl et al.. After investigation, we reveal several weaknesses of these schemes. First, Bindu et al.¡s protocol suffers from an insider impersonation attack if a malicious user obtains a lost smart card. Second, both Goriparthi et al.¡s and Wang et al.¡s protocols cannot withstand a DoS attack in the password change phase, i.e. an attacker can involve the phase to make user¡s password never be used in subsequent authentications. Third, Holbl et al.¡s protocol is vulnerable to an insider attack since a legal but malevolent user can deduce KGC¡s secret key.**

Keywords- password authentication protocol; insider attack; denial-of-service attack; smart card lost problem; mutual authentication; man-in-the-middle attack

I. INTRODUCTION

Authentication protocols provide two entities to ensure that the counterparty is the intended one whom he attempts to communicate with over an insecure network. These protocols can be considered from three dimensions: type, efficiency and security.

In general, there are two types of authentication protocols, the password-based and the public-key based. In a password-based protocol, a user registers his account and password to a remote server. Later, he can access the remote server if he can prove his knowledge of the password. The server usually maintains a password or verification table but this will make the system easily subjected to a stolen-verifier attack. To address this problem, recent studies suggest an approach without any password or verification table in the server. Moreover, to enhance password protection, recent studies also introduce a tamper-resistant smart card in the user end. In a public key-based system, a user should register himself to a trust party, named KGC (Key Generation Center) to obtain his public key and corresponding private key. Then, they can be recognized by a network entity through his public key. To simplify the key management, an identity-based public-key cryptosystem is usually adopted, in which KGC issues user¡s ID as public key and computes corresponding private key for a user.

Considering computational efficiency in an authentication protocol, researchers employs low computational techniques

such as secure one-way hash functions or symmetric key encryptions rather than much expensive computation like asymmetric key encryptions (i.e., RSA, ECC, ElGamal, and bilinear pairings). As considering communication efficiency, it usually to reduce the number of passes (rounds) of a protocol since the round efficiency is more significant than the computation efficiency.

The most important dimension of an authentication protocol is its security, and it should ensure secure communications for any two legal entities over an insecure network. Attackers easily eavesdrop, modify or intercept the communication messages on the open network. Hence, an authentication protocol should withstand various attacks, such as password guessing attack, replay attack, impersonation attack, insider attack, and man-in-the-middle attack.

In recent decade, many secure authentication protocols [1-41] were proposed. In 2008, Bindu et al. [14] proposed an improvement from Chien and Chen¡s work [3]. Their protocol is a smart-card based password authentication protocol and employs symmetric key cryptosystem. They claimed that their protocol is secure, provides user anonymity, and prevent from various attacks: replay attack, stolen-verifier attack, password guessing attack, insider attack, and man-in-the-middle attack. In 2009, Goriparthi et al. proposed a scheme [27] based on Das et al.¡s protocol [2] and can avoid the weakness existing in Chou et al.¡s [5]. Goriparthi et al.¡s protocol is also a smart card based password authentication protocol and bases on bilinear pairings. They claimed that their protocol is secure and can withstand replay attack and insider attack. In the same year, Wang et al. [31] also proposed an improvement based on Das et al.¡s protocol [2]. Their scheme is a smart card based password authentication protocol as well and uses secure one-way hash function. Also in 2009, Holbl et al. [40] improved from two identity-based authentication protocols, Hsieh et al. [1] and Tseng et al. [8]. Their protocols are neither password-based nor smart card based protocols. They employ identity-based ElGamal cryptosystem. Although all of the above schemes claimed that they are secure; however, in this paper, we will demonstrate some security vulnerabilities of these protocol in Bindu et al.¡s [14], Goriparthi et al.¡s [27], Wang et al.¡s [31], and Holbl et al.¡s work, correspondingly.

II.   REVIEW AND ATTACK ON BINDU ET AL.'S PROTOCOL

In this section, we first review Bindu et al.¡s protocol[14] and then show an insider attack launched by an insider who is supposed to have obtained another legal user¡s smart card.

A.   Review

There are three phases in Bindu et al.¡s protocol: the registration phase, the login phase, and the authentication phase.

In the registration phase, server S issues to user i a smart card which contains m_i and I_i, where m_i=H$(ID_i \oplus s) \oplus$ H$(s) \oplus$ H(PW_i), I_i=H$(ID_i \oplus s) \oplus s$, and s is S¡s secret key.

When i wants to login to S, he starts the login phase and computes $r_i = g^x$ (x is a random number chosen by i), $M = m_i \oplus$ H(PW_i), $U = M \oplus r_i$, $R = I_i \oplus r_i =$ H$(ID_i \oplus s) \oplus s \oplus r_i$, and E$_R[r_i, ID_i, T]$ (T is a timestamp, and E$_R[r_i, ID_i, T]$ is a ciphertext encrypted by the secret key R). He then sends $\{U, T,$ E$_R[r_i, ID_i, T]\}$ to S.

In the authentication phase, after receiving $\{U, T,$ E$_R[r_i, ID_i, T]\}$ at time T_s, S computes $R = U \oplus$ H$(s) \oplus s = M \oplus r_i \oplus$ H$(s) \oplus s$ $= m_i \oplus$ H$(PW_i) \oplus r_i \oplus$ H$(s) \oplus s =$ H$(ID_i \oplus s) \oplus$ H$(s) \oplus$ H$(PW_i) \oplus$ H$(PW_i) \oplus r_i \oplus$ H$(s) \oplus s =$ H$(ID_i \oplus s) \oplus r_i \oplus s$, decrypts E$_R[r_i, ID_i, T]$, checks to see if $T_s - T$ is less than ΔT, and compares R with H$(ID_i \oplus s) \oplus s \oplus r_i$ to see if they are equal. If they are, he sends $\{T_s,$ E$_R[r_s, r_i+1, T_s]\}$ to i, where $r_s = g^y$ and y is a random number chosen by S. After that, i verifies the validity of the timestamp T_s, decrypts E$_R[r_s, r_i+1, T_s]$, and checks to see if r_i+1 is correct or not. If it is, S is authentic. Then, i sends $\{$E$_{Kus}[r_s+1]\}$ to S, where $K_{us} = r_s^x = g^{xy}$. Finally, S decrypts the received message $\{$E$_{Kus}[r_s+1]\}$ and checks to see if the value of r_s+1 is correct or not. If it is, i is authentic.

B.   Attack

If C lost his smart card and the card is got by an insider E, E can impersonate C to log into S. We show the attack in the following.

For that C¡s smart card stores m_c=H$(ID_c \oplus s) \oplus$ H$(s) \oplus$ H(PW_c) and I_c=H$(ID_c \oplus s) \oplus s$, and E¡s smart card stores m_e=H$(ID_e \oplus s) \oplus$ H$(s) \oplus$ H(PW_e) and I_e=H$(ID_e \oplus s) \oplus s$, suppose E gets C¡s smart card but doesn¡t have the knowledge of PW_c, E can choose a random number x and computes $r_c = g^x$, $V = m_e \oplus I_e \oplus$ H(PW_e)=H$(s) \oplus s$, $M = I_c \oplus V =$ H$(ID_c \oplus s) \oplus s \oplus$ H$(s) \oplus s$ =H$(ID_c \oplus s) \oplus$ H(s) which equals $m_c \oplus$ H(PW_c), $U = M \oplus r_c$, and $R = I_c \oplus r_c$. Then, E masquerades as C by sending $\{U, T,$ E$_R[r_c, ID_c, T]\}$ to S. After receiving the message, S computes $R = U \oplus$ H$(s) \oplus s$ and compares R with H$(ID_c \oplus s) \oplus s \oplus r_c$. If they are equal, S sends C the message $\{T_s,$ E$_R[r_s, r_c+1, T_s]\}$. E intercepts the message, decrypts E$_R[r_s, r_c+1, T_s]$, and uses r_s to compute $K_{us} = r_s^x = g^{xy}$. E then can send a correct message $\{$E$_{Kus}[r_s+1]\}$ to S, to let S authenticate him as C. In other words, insider E can successfully launch an insider attack if the user¡s smart card is lost.

More clarity, we demonstrate why $R = U \oplus$ H$(s) \oplus s$ is equal to H$(ID_c \oplus s) \oplus s \oplus r_c$ by the following equations.

$R = U \oplus$ H$(s) \oplus s$
$= M \oplus r_c \oplus$ H$(s) \oplus s$ ($\because U = M \oplus r_c$)
$= I_c \oplus V \oplus r_c \oplus$ H$(s) \oplus s$ ($\because M = I_c \oplus V$)
$=$ H$(ID_c \oplus s) \oplus s \oplus V \oplus r_c \oplus$ H$(s) \oplus s$ ($\because I_c$=H$(ID_c \oplus s) \oplus s$)
$=$ H$(ID_c \oplus s) \oplus s \oplus$ H$(s) \oplus s \oplus r_c \oplus$ H$(s) \oplus s$ ($\because V$=H$(s) \oplus s$)
$=$ H$(ID_c \oplus s) \oplus s \oplus r_c$

III.   REVIEW AND ATTACK ON GORIPARTHI ET AL.'S PROTOCOL

In this section, we first review Goriparthi et al.¡s scheme [27] and then demonstrate a DoS attack on the password change phase of the protocol, which will make user¡s password never be used in subsequent authentications.

A.   Review

In the password change phase of Goriparthi et al.¡s protocol, when client C wants to change his password PW, he keys his ID and PW to his smart card. According their protocol, the smart card only checks ID while no mechanism to verify the validity of PW. If the ID is matched with the one stored in the smart card, the smart card will continuously ask C a new password $PW*$, and then compute $Reg^*_{ID} = Reg_{ID}$ ¡ h(PW) + h$(PW*)$ = s¡h(ID) + h$(PW*)$, where Reg_{ID} = s¡h(ID) + h(PW) is issued by the server and stored in C¡s smart card in the registration phase, h$(¡)$ is a map-to-point hash function, h$:\{0,1\}* \to G_1$, and G_1 is a group on an elliptic curve. Finally, the smart card will replace Reg_{ID} with Reg^*_{ID}.

B.   Attack

In the protocol, assume that an attacker temporarily gets C¡s smart card. He arbitrarily selects two passwords PW' and PW'' as the old and the new ones, respectively. The smart card will then compute $Reg'_{ID} = Reg_{ID}$ ¡ h(PW') + h(PW'') = s¡h(ID) + h(PW) ¡ h(PW') + h(PW'') and replace Reg_{ID} with Reg'_{ID}. This will make C¡s original password PW never be used in subsequent authentications and thus cause denial of service.

IV.   REVIEW AND ATTACK ON THE PROTOCOL OF WANG ET AL.¡S PROTOCOL

In this section, we first review Wang et al.¡s protocol [31] and then show the protocol has the same weakness ¡ it suffers a DOS attack in password change phase ¡ like Goriparthi et al.¡s work [27].

A.   Review

In Wang et al.¡s protocol , C inserts his smart card, keys PW, and requests to change the password PW to a new one $PW*$. On receiving the request, the smart card computes $N_i* = N_i \oplus$ H$(PW) \oplus$ H$(PW*)$ and replaces N_i with N_i*, where $N_i =$ H$(PW_i) \oplus$ H(x) is stored in C¡s smart card, PW_i is chosen by

the user when he registers himself to the remote server S, and x is S¡s secret key..

B. Attack

Obviously, this protocol also exits the same weakness like Goriparthi et al.¡s work [27]. Since if an attacker temporarily gets C¡s smart card, he can use two arbitrary values PW' and PW'' to ask the smart card to update its storage through password change protocol. The smart card will compute $N_i' = N_i \oplus H(PW') \oplus H(PW'')$ and replace N_i with N_i'. From then on, client C can never pass the subsequent authentications.

V. REVIEW AND ATTACK ON THE PROTOCOL OF HOLBL ET AL.'S PROTOCOL

Holbl et al. [40] proposed two improvements of two-party key agreement and authentication protocols. In the following, we first briefly review their schemes and then present their weaknesses.

A. Review of Holbl et al.¡s First Protocol

Holbl et al.¡s first protocol consists of three phases: the system setup phase, the private key extraction phase, and the key agreement phase.

In the system setup phase, KGC chooses a random number x_s and keeps it secret. He computes $y_s = g^{x_s}$ as public key.

In the private key extraction phase, for each user who has identity ID_i, KGC selects a random number k_i, and calculates his private key $v_i = I_i k_i + x_s u_i \pmod{p_¡1}$ and corresponding public key $u_i = g^{k_i} \pmod{p}$, where $I_i = H(ID_i)$.

In the key agreement phase, user A chooses a random number r_a, computes $t_a = g^{r_a}$, and then sends $\{u_a, t_a, ID_a\}$ to user B. After receiving $\{u_a, t_a, ID_a\}$, B chooses a random number r_b, calculates $t_b = g^{r_b}$, and then sends $\{u_b, t_b, ID_b\}$ back to A. Finally, A and B can respectively compute their common session key, $K_{AB} = (u_b^{I_b} y_s^{u_b} t_b)^{(v_a+r_a)} = g^{(v_b+r_b)\cdot(v_a+r_a)}$ and $K_{BA} = (u_a^{I_a} y_s^{u_a} t_a)^{(v_b+r_b)} = g^{(v_a+r_a)\cdot(v_b+r_b)}$, where $I_a = H(ID_a)$ and $I_b = H(ID_b)$.

B. Attack on Holbl et al.¡s first protocol

Assume that an insider C calculates $I_c = H(ID_c)$ and $q = \gcd(I_c, u_c)$, and computes $w = I_c/q$, $z = u_c/q$, and $j = v_c/q$, where v_c is C¡s private key. Hence, $\gcd(w, z) = 1$. Then, he can use the extended Euclid¡s algorithm to find α and β both satisfying that $\alpha_¡w + \beta_¡z = 1$. As a result, he can obtain both x_s and k_c, since $v_c = 1 \, j_¡c_¡q_c = (\alpha_¡w + \beta_¡z)_¡j_¡q_c = (\alpha_¡I_c/q + \beta_¡u_c/q)_¡j_¡q = (\alpha_¡I_c + \beta_¡u_c)_¡j = I_c¡(\alpha_¡j) + (\beta_¡j)_¡u_c$ and $v_c = I_c¡k_c + x_s¡u_c$, where x_s is KGC¡s secret key and k_c is a random number selected by KGC satisfying $u_c = g^{k_c}$. More clearly, the value x_s he obtains is equal to $\beta_¡j$.

After obtaining x_s, C can deduce any user¡s private key in the same manner. As an example, in the following, we demonstrate how C can deduces user i¡s private key, k_i. C calculates $I_i = H(ID_i)$ and $q_i = \gcd(I_i, u_i)$, computes $w_i = I_i/q_i$ and $z_i = u_i/q_i$, and then uses the extended Euclid¡s algorithm to compute γ and ε satisfying that $\gamma_¡w_i + \varepsilon_¡z_i = 1$. Finally, since v_i

$= 1 \, j_i¡q_i = (\gamma_¡w_i + \varepsilon_¡z_i)_¡j_i¡q_i = (\gamma_¡I_i/q_i + \varepsilon_¡u_i/q_i)_¡j_i¡q_i = (\gamma_¡I_i + \varepsilon_¡u_i)_¡j_i = I_i¡(\gamma_¡j_i) + (\varepsilon_¡j_i)_¡u_i$ and $v_i = I_i¡k_i + x_s¡u_i$, he can calculate $j_i = x_s/\varepsilon$ and thus obtains i¡s private key by computing $v_i = j_i¡q_i$. With the knowledge of i¡s private key, insider C can impersonate user i to communicate with any other legal user.

C. Review of Holbl et al.¡s second protocol

Holbl et al.¡s second protocol consists of three phases: the system setup phase, the private key extraction phase, and the key agreement phase.

The system setup phase of this protocol is the same as the one in the first protocol.

In the private key extraction phase, with each user having his identity ID, KGC selects a random number k_i, and calculates i¡s private key $v_i = k_i + x_s¡H(ID_i, u_i)$ and public key $u_i = g^{k_i}$.

In the key agreement phase, user A chooses a random number r_a, computes $t_a = g^{r_a}$, and then sends $\{u_a, t_a, ID_a\}$ to user B. After receiving $\{u_a, t_a, ID_a\}$, B chooses a random number r_b, calculates $t_b = g^{r_b}$, and then sends $\{u_b, t_b, ID_b\}$ to A. Finally, A and B can compute their common session key, $K_{AB} = (u_b¡y_s^{H(ID_b,u_b)}\cdot t_b)^{(v_a+r_a)} = g^{(v_b+r_b)(v_a+r_a)}$ and $K_{BA} = (u_a¡y_s^{H(ID_a,u_a)}¡t_a)^{(v_b+r_b)} = g^{(v_a+r_a)(v_b+r_b)}$, respectively.

D. Attack on Holbl et al.¡s second protocol

Likewise, we can launch the same attack, as do in the first one, on this scheme. Since $\gcd(1, H(ID_c, u_c)) = 1$, an insider C can use the extended Euclid¡s algorithm to find α and β both satisfying that $\alpha_¡1 + \beta_¡H(ID_c, u_c) = 1$. And since $v_c = k_c + x_s¡H(ID_c, u_c)$ and $1 = (k_c/v_c)_¡1 + (x_s/v_c)_¡H(ID_c, u_c)$, he can obtain both x_s and k_c by letting $x_s = \beta_¡v_c$ and $k_c = \alpha_¡v_c$, where v_c is C¡s private key, x_s is KGC¡s secret key and k_c is a random number selected by KGC satisfying $u_c = g^{k_c}$. Consequently, similar to the result as shown in the attack of the first protocol, insider C can impersonate user i to communicate with any other legal user.

VI. CONCLUSION

In the paper we have investigate four authentication protocols. In Bindu et al.¡s scheme [14], an insider can employ his own secrecy in the smart card issued from the server to successfully impersonate another user by getting the victim¡s smart card. In both Goriparthi et al.¡s and Wang et al.¡s schemes, their password change phases are easily subjected to a DOS attack, because no proper mechanism to verify user¡s input password. Finally, in Holbl et al.¡s scheme, any legal user can extract KGC¡s private key.

REFERENCES

[1] B. T. Hsieh, H. M. Sun, T. Hwang, C. T. Lin, ¡An Improvement of Saeednia¡s Identity-based Key Exchange Protocol¡, Information Security Conference 2002, pp. 41-43, 2002.

[2] M. L. Das, A. Saxena, V. P. Gulati, ¡A dynamic ID-based remote user authentication scheme¡, *IEEE Transactions on Consumer Electronics*, Vol. 50, No. 2, pp. 629-631, May 2004.

[3] H. Y. Chien, C. H. Chen, ¡A Remote Password Authentication Preserving User Anonymity,¡ *Proceedings of the 19th International Conference on Advanced Information Networking and Applications* (AINA ¡05), Vol.2, pp. 245-248, March 2005.

[4] J. S. Chou, M. D. Yang, G. C. Lee, ¡Cryptanalysis and improvement of Yang-Wang password authentication schemes¡, http://eprint.iacr.org/2005/466, December 2005.

[5] J.S. Chou, Y. Chen, J. Y. Lin, ¡Improvement of Das et al.'s remote user authentication scheme¡, http://eprint.iacr.org/2005/450.pdf, December 2005.

[6] M. Peyravian, C. Jeffries, ¡Secure remote user access over insecure networks¡, *Computer Communications*, Vol. 29, No. 5, pp. 660-667, March 2006.

[7] I. E. Liao, C. C. Lee, M. S. Hwang, ¡A password authentication scheme over insecure networks¡, *Journal of Computer and System Sciences*, Vol. 72, No. 4, pp. 727-740, June 2006.

[8] Y. M. Tseng, ¡An Efficient Two-Party Identity-Based Key Exchange Protocol¡, *Informatica*, Vol. 18, No. 1, pp. 125-136, January 2007.

[9] J. Nam, Y. Lee, S. Kim, D. Won, ¡Security weakness in a three-party pairing-based protocol for password authenticated key exchange¡, *Information Sciences*, Vol. 177, No. 6, pp. 1364-1375, March 2007.

[10] H. R. Chung, W. C. Ku, ¡Three weaknesses in a simple three-party key exchange protocol¡, *Information Sciences*, Vol. 178, No. 1-2, pp. 220-229, January 2008.

[11] T. H. Chen, W. B. Lee, ¡A new method for using hash functions to solve remote user authentication¡, Computers *& Electrical Engineering*, Vol. 34, No. 1, pp. 53-62, January 2008.

[12] H. B. Chen, T. H. Chen, W. B. Lee, C. C. Chang, ¡Security enhancement for a three-party encrypted key exchange protocol against undetectable on-line password guessing attacks¡, *Computer Standards & Interfaces*, Vol. 30, No. 1-2, pp. 95-99, January 2008.

[13] H. Guo, Z. Li, Y. Mu, X. Zhang, ¡Cryptanalysis of simple three-party key exchange protocol¡, *Computers & Security*, Vol. 27, No. 1-2, pp. 16-21, March 2008.

[14] C. S. Bindu, P. C. S. Reddy, B. Satyanarayana, ¡Improved remote user authentication scheme preserving user anonymity¡, *International Journal of Computer Science and Network Security*, Vol. 8, No. 3, pp. 62-65, March 2008.

[15] Y. Lee, J. Nam, D. Won, ¡Vulnerabilities in a remote agent authentication scheme using smart cards¡, *LNCS: AMSTA*, Vol. 4953, pp. 850-857, April 2008.

[16] W. S. Juang, S. T. Chen, H. T. Liaw, ¡Robust and efficient password-authenticated key agreement using smart cards¡, *IEEE Transactions on Industrial Electronics*, Vol. 55, No. 6, pp. 2551-2556, June2008.

[17] W. S. Juang, W. K. Nien, ¡Efficient password authenticated key agreement using bilinear pairings¡, *Mathematical and Computer Modelling*, Vol. 47, No. 11-12, pp. 1238-1245, June 2008.

[18] J. Y. Liu, A. M. Zhou, M. X. Gao, ¡A new mutual authentication scheme based on nonce and smart cards¡, *Computer Communications*, Vol. 31, No. 10, pp. 2205-2209, June 2008.

[19] M. Holbl, T. Welzer, B. Brumen, ¡Improvement of the Peyravian-Jeffries¡s user authentication protocol and password change protocol¡, *Computer Communications*, Vol. 31, No. 10, pp. 1945-1951, June 2008.

[20] J. L. Tsai, ¡Impersonation attacks on Rhee et al.¡s authentication scheme¡, *http://dtim.mis.hfu.edu.tw/2008/paper/C044.pdf*, June 2008.

[21] J. L. Tsai, ¡Efficient multi-server authentication scheme based on one-way hash function without verification table¡, *Computers & Security*, Vol. 27, No. 3-4, pp. 115-121, May-June 2008.

[22] E. J. Yoon, K. Y. Yoo, ¡Improving the novel three-party encrypted key exchange protocol¡, *Computer Standards & Interfaces*, Vol. 30, No. 5, pp. 309-314, July 2008.

[23] R. C. Phan, W. C. Yau, B. M. Goi, ¡Cryptanalysis of simple three-party key exchange protocol (S-3PAKE)¡, *Information Sciences*, Vol. 178, No. 13, pp. 2849-2856, July 2008.

[24] C. C. Chang, J. S. Lee, T. F. Cheng, ¡Security design for three-party encrypted key exchange protocol using smart cards¡, ACM Proceedings of the 2nd international conference on Ubiquitous information management and communication, pp. 329-333, 2008.

[25] T. Xiang, K. Wong, X. Liao, ¡Cryptanalysis of a password authentication scheme over insecure networks¡, *Computer and System Sciences*, Vol. 74, No. 5, pp. 657-661, August 2008.

[26] G. Yang, D. S. Wong, H. Wang, X. Deng, ¡Two-factor mutual authentication based on smart cards and passwords¡, *Journal of Computer and System Sciences*, Vol. 74, No. 7, pp.1160-1172, November 2008.

[27] T. Goriparthi, M. L. Das, A. Saxena, ¡An improved bilinear pairing based remote user authentication scheme¡, *Computer Standards & Interfaces*, Vol. 31, No. 1, pp. 181-185, January 2009.

[28] H. S. Rhee, J. O. Kwon, D. H. Lee, ¡A remote user authentication scheme without using smart cards¡, *Computer Standards & Interfaces*, Vol. 31, No. 1, pp. 6-13, January 2009.

[29] Y. Liao, S. S. Wang, ¡A secure dynamic ID based remote user authentication scheme for multi-server environment¡, *Computer Standards & Interfaces*, Vol. 31, No. 1, pp. 24-29, January 2009.

[30] J. Munilla, A. Peinado, ¡Security flaw of Holbl et al.¡s protocol¡, *Computer Communications*, Vol. 32, No. 4, pp.736-739, March 2009.

[31] Y. Y. Wang, J. Y. Liu, F. X. Xiao, J. Dan, ¡A more efficient and secure dynamic ID-based remote user authentication scheme¡, *Computer Communications*, Vol. 32, No. 4, pp. 583-585, March 2009.

[32] H. C. Hsiang, W. K. Shih, ¡Weaknesses and improvements of the Yoon¡Ryu¡Yoo remote user authentication scheme using smart cards¡, *Computer Communications*, Vol. 32, No. 4, pp. 649-652, March 2009.

[33] D. Z. Sun, J. P. Huai, J. Z. Sun, J. X. Li, ¡Cryptanalysis of a mutual authentication scheme based on nonce and smart cards¡, *Computer Communications*, Vol. 32, No. 6, pp. 1015-1017, April 2009.

[34] S. K. Kim , M. G. Chung, ¡More secure remote user authentication scheme¡, *Computer Communications*, Vol. 32, No. 6, pp. 1018-1021, April 2009.

[35] H. R. Chung, W. C. Ku, M. J. Tsaur, ¡Weaknesses and improvement of Wang et al.'s remote user password authentication scheme for resource-limited environments¡ , *Computer Standards & Interfaces*, Vol. 31, No. 4, pp. 863-868, June 2009.

[36] J. Xu, W. T. Zhu, D. G. Feng, ¡An improved smart card based password authentication scheme with provable security¡ , *Computer Standards & Interfaces*, Vol. 31, No. 4, pp. 723-728, June 2009.

[37] J. H. Yang, C. C. Chang, ¡ An ID-based remote mutual authentication with key agreement scheme for mobile devices on elliptic curve cryptosystem¡ , *Computers & Security*, Vol. 28, No. 3-4, pp. 138-143, May-June 2009.

[38] M. S. Hwang, S. K. Chong, T. Y. Chen, ¡DoS-resistant ID-based password authentication scheme using smart cards¡ , *Journal of Systems and Software*, In Press, Available online 12 August 2009.

[39] H.C. Hsiang, W.K. Shih, ¡Improvement of the secure dynamic ID based remote user authentication scheme for multi-server environment¡, *Computer Standards & Interfaces*, Vol. 31, No. 6, pp. 1118-1123, November 2009.

[40] M. Holbl, T. Welzer, ¡Two improved two-party identity-based authenticated key agreement protocols¡ , *Computer Standards & Interfaces*, Vol. 31, No. 6, pp. 1056-1060, November 2009.

[41] C. T. Li, M. S. Hwang, ¡An efficient biometrics-based remote user authentication scheme using smart cards¡, *Journal of Network and Computer Applications*, Vol. 33, No. 1, pp. 1-5, January 2010.

AUTHORS PROFILE

Yalin Chen   received her bachelor degree in the depart. of computer science and information engineering from Tamkang Univ. in Taipei, Taiwan and her MBA degree in the department of information management from National Sun-Yat-Sen Univ. (NYSU) in Kaohsiung, Taiwan. She is now a Ph.D. candidate of the Institute of Info. Systems and Applications of National Tsing-Hua Univ.(NTHU) in Hsinchu, Taiwan. Her primary research interests are data security and privacy, protocol security, authentication, key agreement, electronic commerce, and wireless communication security.

Jue-Sam Chou   received his Ph.D. degree in the department of computer science and information engineering from National Chiao Tung Univ. (NCTU) in Hsinchu, Taiwan,ROC. He is an associate professor and teaches at the department of Info. Management of Nanhua Univ. in Chiayi, Taiwan. His primary research interests are electronic commerce,   data security and privacy, protocol security, authentication, key agreement, cryptographic protocols, E-commerce protocols, and so on.

Chun-Hui Huang   is now a graduate student at the department of Info. Management of Nanhua Univ. in Chiayi, Taiwan. She is also a teacher at Nantou County Shuang Long Elementary School in Nantou, Taiwan. Her primary interests are data security and privacy, protocol security, authentication, key agreement.

Software Metrics: Some degree of Software Measurement and Analysis

Rakesh.L
Department of Computer-Science
SCT Institute of Technology
Bangalore - 560075, India
rakeshsct@yahoo.co.in

Dr.Manoranjan Kumar Singh
PG Dept of Mathematics
Magadh University
Bodhgaya- 824234, India
drmksingh_gaya@yahoo.com

Dr.Gunaseelan Devaraj
Department of Information Technology
Ibri College of Technology
Sultanate of Oman- 516
dgseela@yahoo.com

Abstract— **Measurement lies at the heart of many systems that govern our lives. Measurement is essential to our daily life and measuring has become a common place and well accepted. Engineering discipline use methods that are based on models and theories. Methodological improvements alone do not make an engineering discipline. Measurement encourages us to improve our processes and products. This paper examines the realm of software engineering to see why measurement is needed and also set the scene for new perspective on software reliability metrics and its improvement. Software measurement is not a mainstream topic within software engineering rather it is a diverse collection of fringe topics. Unlike other engineering discipline measurement must become an integral part of software engineering practice.**

Keywords- External Attribute, Reliability model, Fault tolerance.

I. INTRODUCTION

Measurement is a process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them accordingly to clearly defined rules. Software engineering describes the collection of techniques that apply an engineering approach to the development and support of products. Software engineering activities include specifying, analysing, designing, implementing, testing and maintaining. By engineering approach we mean that each activity is understood and controlled, so that there are few surprises as the software is specified, designed, built, and maintained. Whereas computer science provides the theoretical foundations for building for building the software, software engineering focuses on implementing the software in controlled and scientific way. The importance of software engineering cannot be understated, since software pervades our lives. From oven control to airbags, from banking transactions to air traffic control, and from sophisticated power plants to sophisticated weapons, our lives and quality of life depend on software. Such a young discipline has done an admirable job or providing safe, useful, and reliable functionality. But there is a room for great deal of improvement. Engineering discipline use techniques that are based on models which are theoretical in nature. For example, in designing electrical circuits we appeal to theories like Ohms law, which describes the relation between resistance, current

and voltage in the circuit. But the laws of electrical behaviour have evolved by using the scientific method, stating a hypothesis, designing and running an experiment to test its truth and analysing the results. Underpinning the scientific process is measurement, measuring the variables to differentiate cases, measuring the changes in behaviour, and measuring the causes and effects. Once the scientific method suggests the validity of the model or the truth of a theory, we continue to use measurement to apply the theory to practice. It is difficult to imagine electrical, mechanical and civil engineering without a central role of measurement. Indeed, science and engineering can be neither effective nor practical without measurement. But measurement has been considered a luxury in software engineering. For most development projects we fail to set measurable targets for our software products. For example, we promise that the product will be user friendly, reliable and maintainable without specifying clearly and objectively what these terms mean. As a result when the project is complete, we cannot tell if we have met our goals. This situation has prompted Tom Gilb to state "projects without clear goals will not achieve their goals clearly". We do not quantify or predict the quality of the products we produce. Thus, we cannot tell a potential user how reliable a product will be in terms of likelihood of failure in a given period of use, or how much work will be needed to port the product to a different machine environment [1]. We allow anecdotal evidence to convince us to try yet another revolutionary new development technology, without doing a carefully controlled study to determine if the technology is efficient and effective. When measurements are made, they are often done infrequently, inconsistently and incompletely. The incompleteness can be frustrating to those who make use of the results. For instance a tester may say that there are on a average 55 faults in every 1000 lines of software code. But we are not always told how these results were obtained, how experiments were designed and executed, which entities were measured and how and what were the realistic error margins. Without this information we remain skeptical and unable to decide whether to apply the results to our own situations. Thus, the lack of measurement in software engineering is compounded by the lack of a rigorous approach. We have set a new perspective on software measurement on solid foundation.

II. MOTIVATION

In mathematical analysis a metric has a very specific meaning .It is a rule used to describe how far apart two points are. More formally, a metric is a function m defined on pairs of objects x and y such that m(x,y) represents the distance between x and y. Such metrics must satisfy certain properties:

- m(x,x) = 0 for all x : that is, the distance from point x to itself is zero;

- m(x,y) = m(y,x) for all x and y: that is, the distance from x to y is same as the distance from y to x;

- m(x,z) ≤ m(x,y)+m(y,z) for all x,y and z: that is, the distance from x to z is no larger than the distance measured by stopping through an intermediate point.

A prediction system consists of a mathematical model together with a set of prediction procedures for determining unknown parameters and interpreting the results. When we talk about measuring something, we usually mean that we wish to assess some entity that already exists. This measurement for assessment is very helpful in understanding what exist now or what happened in the past [2]. However, in many circumstances, we would like to predict an attribute of some entity that does not yet exist. For instance, suppose we are building a software system that must be highly reliable, such as the control software for an aircraft, power plant, or X-ray machine. The software development may take some time and we want to provide early assurance that the system will meet reliability targets. To provide reliability indicators before the system is complete, we can build a model of the factors that affect reliability, and then predict the likely reliability based on our understanding of the system while it is still under development. In general, measurement for prediction always requires some kind of mathematical model that relates the attributes to be predicted to some other attributes that we can measure. The model need not be complex to be useful. Suppose we want to predict the number of pages, M, that will print out as a source code program, so that we can order sufficient paper or estimate the time it will take to do the printing. We can use the very simple model,

$$M = x/a \qquad (1)$$

Where x is a variable representing a measure of source code program length in lines of code, and a is constant representing the average number of lines per page. Effort predication is essential and universally needed. A common generic model for predicting the effort required in software projects has the form,

$$E = aS^b \qquad (2)$$

Where E is effort in person months, S is the size in lines of code of the system to be developed, and a and b are constants. There are numerous examples can be represented as "mathematical" metrics for software [3]. We can hope that every program satisfies its specification completely, but this is rarely the case. Thus, we can view program correctness as a measure of the extent to which a program satisfies its specification, and define a metric m(S,P) where the entities S(Specification) and P(Program) are both products. Let us elaborate this idea to classical fault tolerant technique like N-Version programming which is quite popular in safety critical systems. The approach involves developing N different versions of critical software components independently. Theoretically, by having each of the N-different teams solving the same problem without knowledge of what the other teams are doing , the probability that all the teams , or even majority, will make same error is kept small. When the behaviour of the different versions differs, a voting procedure accepts the behaviour of the majority systems. The assumptions, then is that the correct behaviour will always be chosen. However, there may be problems in assuring genuine design independence, so we may be interested in measuring the level of diversity between two designs, or algorithms or programs. We can define a metric m, where m(P$_1$, P$_2$) measures the diversity between two programs P$_1$ and P$_2$. In this case, the entities being measured are products. We can use a similar metric to measure the level of diversity between two methods applied during design, we would be measuring attributes of process entities. To reconcile these mathematically precise metrics with framework we have Proposed, we can consider pairs of entities as a single entity. For instance, having produced two programs satisfying the same specification, we consider the pair of programs to be a single product system, itself having a level of diversity. This approach is consistent with systems view of N-version programming. Where we have implemented N versions of a program, the diversity of the system may be viewed as an indirect measure of the pair wise program diversity. Many attributes of interest in software engineering are not directly measurable. This situation forces to use vectors of measures, with rules of combining the vector elements into a larger, indirect measure [4]. That is indirect measure is defined to be combination of other measures, both direct and indirect. An indirect measure of testing efficiency T is D/E, where D is the number of defects discovered and E is effort in person months. Here D is an absolute scale measure, while E is on the ratio scale measure. Since absolute is stronger than ratio scale, it follows that T is a ratio scale measure. Consequently the acceptable rescalings of T arise from rescaling of E into other measures of effort like person days, person years etc. Many of the measures we have used in our examples are assessment measures. But indirect measures proliferate as prediction measures also.

III. SOFTWARE FOUNDATION

Metrics are standards that are commonly accepted scales that define measurable attributes of entities, their units and their scope. Measurement is process by which numbers or symbols are assigned to attributes of entities (objects) in the real world in such a way as to ascribe them according to define rules. Measure is a relation between an attribute and a measurement scale. In any measurement activity, there are rules to be followed. The rules help us to be consistent in our measurement, as well as providing a basis for interpreting data. Measurement theory tells us the rules, laying the ground work for developing and reasoning about all kinds of measurement. This rule based approach is common in many sciences. For example recall that mathematicians learned about the world by defining axioms of geometry. Then by combining axioms and using their results to support or refute their observations, they expanded their understanding and the set of rules that govern the behaviour of objects. In the same way, we can use rules about measurement to codify our initial understanding, and expand our horizons as we analyse our software. The representational theory is depicted in Figure 1.
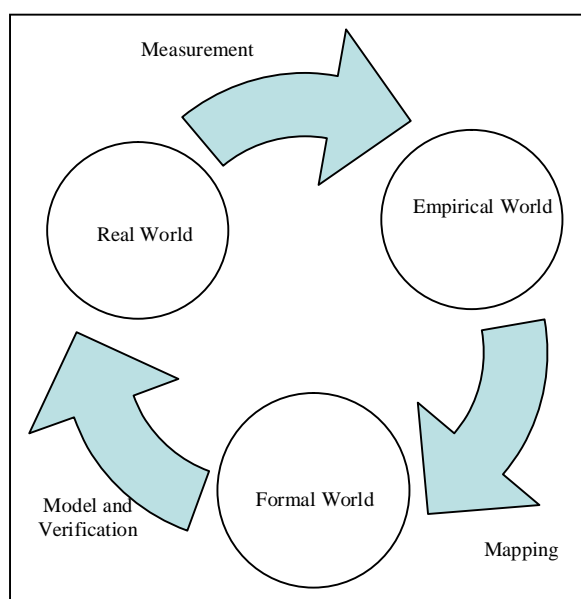
Fig. 1 Mathematical Logical Model

The representational theory of measurement seeks to formalize our intuition about the way the world works. That is, the data we obtain as measures should represent attributes of the entities we observe, and manipulation of the data should preserve relationships that we observe among the entities. Thus, our intuition is the starting point for all our measurement. Formally, we define measurement as a mapping from the empirical world to the formal, relational world. Consequently, a measure is the number or symbol assigned to an entity by this mapping in order to characterize an attribute.

An entity in a software measurement can be any artifacts, design and development activity, verification, quality assurance and resources [5]. An attribute is a feature property of an entity for instance blood pressure of person, cost of a Journey, duration of the software specification process. There are two general types of attributes internal attributes and external attributes. Internal attributes are measured only based on the entity itself. If we say entity as a code, the internal attributes are size, modularity and coupling. External attributes can be measured with respect to how the entity relates to its environment. If we take for instance code as an entity its external attributes are reliability and maintainability. The first obligation of any software measurement activity is identifying the entities and attributes we wish to measure. In software there are three such classes process, products and resources. Processes are collections of software related activities. Products are any artifacts, deliverables or documents that result from a process activity. Resources are entities required by a process activity. The principal objective of software engineering is to improve the quality of software products. But quality, like beauty, is very much in the eyes of beholder. In the philosophical debate about the meaning of software quality, proposed definitions include fitness for purpose, conformance to specification, degree of excellence and timeliness. However, from the measurement perspective, we must be able to define quality in terms of specific software product attribute of interest to the user. That is, we want to know how to measure the extent to which these attributes are present in our software products. This knowledge will enable us to specify and sets target for quality attributes in measurable form. For many years, the user community sought a single model for depicting and expressing quality. The advantage of universal model is clear, it makes easier the comparison of one product with another. In 1992 McCall model was proposed as the basis for an international standard for software quality measurement called, Software product evaluation, Quality characteristics and Guidelines for their use, the standard is more commonly referenced by its assigned standard number, ISO 9126. In this standard software quality is defined to be, "The totality of features and characteristics of a software product that bear on its ability to satisfy stated or implied needs". Then quality is decomposed into six factors. Functionality, reliability, efficiency, usability, maintainability, portability. The standard claims that these six are comprehensive that is any component of software quality can be described in terms of some aspect of one or more of the six factors. In turn each of the six is defined as set of attributes that bear on as relevant aspect of software and each can be refined through multiple levels of sub characteristics. Most of the software quality models identified software reliability as a key high level attribute. So it is not surprising that software reliability has been the most expensively studied of all the quality attributes. Quantitative methods for its assessment date back to early 1970s, evolving from theory of hardware reliability.

IV. RELIABILITY METRICS

Software reliability is the probability or failure free software operation for a specified period of time in a specified environment. Software reliability is also an important factor affecting system reliability. It differs from hardware reliability in that it reflects the design perfection, rather than manufacturing perfection. The high complexity of software is the major contributing factor of software reliability problems. Measurement in software is still in its infancy, no good quantitative methods have been developed to represent software reliability without excessive limitations. Software reliability is an important attribute of software quality, together with functionality, usability, performance, serviceability, capability, installability, maintainability and documentation. The word reliability is commonly used in everyday life. When a product such as a car or washing machine breaks down, the user is forcibly made aware of the limited reliability of the product [6]. The reliability R of the product can therefore be defined as the probability that the product continues to meet the specification, over a given time period, subject to given environmental conditions. If however, as the time goes on the product fails to meet its specification, then it is considered to have failed. The unreliability F of the product can be defined as the probability that the product fails to meet the specification, over a given period of time. Both reliability and unreliability vary with time. Reliability R(t) decreases with time, an item that has just been tested and shown to meet specification has a reliability of 1 when first placed in service. One year later this may have decreased to 0.5. Unreliability F(t) increases with time, an item that has just been tested and shown to meet specification has an unreliability of 0 when first placed in service, increasing to say 0.5 after one year. Since, at any time t, the product has either survived or failed, the sum of reliability and unreliability must be 1, that is the events are complementary and given by,

$$R(t) + F(t) = 1 \qquad (3)$$

We can now discuss the relationship between quality and reliability. The reliability of a product is its ability to retain its quality as time progresses. Thus a product can only have high quality if it also has high reliability. High initial quality is of little use if it is soon lost. The opposite is, however, not true, a product with high reliability does not necessarily have high quality, but may be merely retaining low quality over a long period of time. When a product is up that is working satisfactorily, it is available for use. When the product is down, i.e. being repaired, it is unavailable for use. It is important to have average measure of the degree to which the product is either available or unavailable. The availability of the product is the fraction of the total test interval that is performing within specification, i.e. up, thus we have,

Availability = Total up time / test Interval
= Total up time/ Total up time + total down time (4)

Unavailability U is similarly defined as the fraction of the total test interval that it is not performing to specification, i.e. failed or down, thus we have,

Unavailability = Total down time/Test Interval (5)

As availability depends on reliability, availability can therefore be increased by increasing mean time between failures (MTBF), i.e. reducing mean failure rate. Availability depend on mean down time, we can increase availability by reducing mean down time, MDT. Thus availability also depends on maintainability, i.e. how quickly the product can be repaired and put back into service. Computers are now widely used in all branches of engineering. Many industrial processes, steel, chemicals, food, gas and electricity generation, rely on computers to monitor and control critical process variables. The reliability of this control is therefore dependent on the reliability of the computer [7]. Furthermore microcomputers form an integral part of a wide range of electronic systems. These embedded microcomputers use computer programs, i.e. software to perform functions previously performed by electronic circuits, i.e. hardware. For example the calculation of a square root can be implemented using either hardware that is the complex electronic circuit or software, a single statement in a high level programming language. Performing functions in software rather than hardware can therefore lead to a simpler, more robust overall system. Since software is vital to the performance of a large number of engineering functions, its reliability should be closely studied. Each copy of a computer program is identical to the original, so that failures due to product variability, often common with hardware, cannot occur with software. Also, unlike, hardware, software cannot usually degrade with time, in the few special cases that degradation does occur it is easy to restore to the original standard [9]. However, software can fail to perform the required function due to undetected errors in the program. If a software error exists in the original program then the same error exists in all copies. If this error produces a failure in a certain set of circumstances, then failure will always occur under these circumstances with possibly serious consequences. Many programs consist of a large number of individual statements and logical paths so that the probability of a significant number of errors being present is high. A software reliability effort is therefore required to minimize the number of errors present by the use of systematic programming methods, checking and testing.

V. RESULTS AND DISCUSSION

Reliability can be quantified in terms of failure probability, failure rate, and mean time between failures. Many branches of engineering are concerned with the development and production of systems which are made up of several simpler elements or components [8].

A. Quantifying Reliability for Series System

In the Figure 2 a system of m elements in series or cascade with individual reliabilities $R_1, R_2, \ldots, R_i, \ldots, R_m$ respectively.
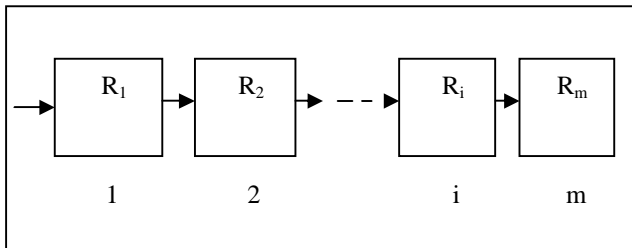
Fig. 2 Reliability of Series System

The system will only survive if every element survives, if one element fails then the system fails. Assuming that the reliability of the other elements, then the probability that the system survives is the probability that element 1 survives and the probability that 2 survives and the probability that element 3 survives, etc. The system reliability R_{SYST} is therefore the product of the individual element reliabilities i.e.,

$$R_{SYST} = R_1 R_2 \ldots R_i \ldots R_m \qquad (5)$$

It is the reliability of the series system. If we further assume that each of the elements can be described by a constant failure rate λ, and λ_i is the failure rate of the ith, element, the failure rate of system of m elements in series is given by,

$$\lambda_{SYST} = \lambda_1 + \lambda_2 + \ldots \lambda_i + \ldots + \lambda_m \qquad (6)$$

This means that the overall failure rate for a series system is the sum of the individual element or component failures rates. From the equation (6), it is evident that, keeping the number of elements in a series system minimum, the system failure rate will be minimum and the reliability is maximum. Protective systems are characterized by having element and system unreliabilities F that are very small. The corresponding element and system reliabilities R are therefore very close to 1, for example, 0.9999 may be typical. In this situation calculation of system reliability may be unwieldy and an alternative equation involving unreliabilities may be more useful. If the individual F_i are small, i.e. $F_i \ll 1$, the terms involving the products of Fs can be neglected giving the

approximate equation for unreliabilities of series system with small F's. The system unreliability is approximately the sum of the element unreliabilites.

B. Quantifying Reliability for Parallel System

The Figure 3 shows an overall system consisting of n individual elements or systems in parallel with individual unreliabilities $F_1, F_2 \ldots F_j, \ldots F_n$ respectively.
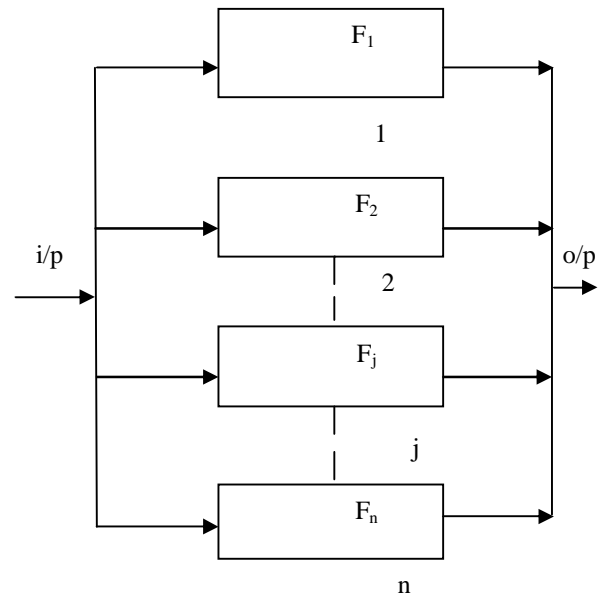
Fig. 3 Reliability of parallel System

All of the elements are expected to be working normally, i.e. to be active, and each one is capable of meeting the functional requirements placed on the overall system. However, only one element/system is necessary to meet these requirements, the remainder increase the reliability of the overall system. This is termed active redundancy. The overall system will only fail if every element or system fails, if one element or system survives the overall system survives. Assuming that the reliability of each element or system is independent of the reliability of the other elements, then the probability that the overall system fails is the probability that element/system 1 fails and the probability that 2 fails and the probability that 3 fails, etc. The overall system unreliability F_{SYST} is therefore the product of the individual element system unreliabilities.

$$F_{SYST} = F_1 F_2 \ldots F_j \ldots F_n \qquad (7)$$

This is the unreliability of parallel system. Comparing, we see

that for series systems, system reliability is the product of element reliabilities, whereas for parallel systems, system unreliability is the product of element unreliabilities. Often the individual elements are identical, so that $F_1 = F_2 = \ldots = F_i = \ldots = F_n = F$, this gives,

$$F_{SYST} = F^n \qquad (8)$$

Thus if $F = 0.1$ and there are four channels in parallel we $F_{SYST} = 10^{-4}$. We see, therefore, that increasing the number of individual elements in parallel system increases the overall system reliability.

C. Reliability model applied to software

In spite of all efforts to ensure that the software is free from errors, some residual errors or bugs often persist [10]. The reliability model presented assumes that the average rate at which software bugs are detected and corrected from similar programs is approximately constant. The software failure rate will then be proportional to the number of remaining bugs. Thus if we assume no new bugs are created during the debugging process and all detected errors are corrected then we have,

Fractional number of residual bugs = Fractional number of total bugs – Fractional number of corrected bugs.i.e.

$$\epsilon_r(\tau) = \{ E_T / I_T \} - \epsilon_c(\tau) \qquad (9)$$

where τ = debugging time in months
E_T = Total number of errors.
I_T = Total number of Instructions

In this model the fractional number of corrected bugs $\epsilon_c(\tau)$ is proportional to τ, i.e.

$$\epsilon_c(\tau) = \rho\tau \qquad (10)$$

where ρ = fractional rate at which errors are removed per month.

If the debugging process is concluded at time τ_0, $\epsilon_c(\tau)$ will therefore remain at the constant value $\epsilon_c(\tau_0)$ and $\epsilon_r(\tau_0)$ will remain at a constant value,

$$\epsilon_r(\tau_0) = \{ E_T / I_T \} - \epsilon_c(\tau_0) \qquad (11)$$

The failure rate λ of the software is then proportional to $\epsilon_r(\tau_0)$, the fractional number of bugs left in the program, i.e.

$$\lambda = K \epsilon_r(\tau_0) \qquad (12)$$

The values of (E_T / I_T), ρ and K must be found by experimental testing before a value of λ can be established.

This model for software reliability is an attempt to quantify reliability for software.

D. Methods of improving software reliability

Specification: Many of the errors recorded during software development originate in the specification. The software specification must describe fully and accurately the requirements of the program. There are no safety margins in software design as in hardware design. The specification must be logically complete, i.e. must cover all possible input conditions and output requirements.

Software systems design: This follows the specification and is often a flow chart which defines the program structure, test points, limits, etc. To be reliable a program must be robust, i.e. it should be able to survive error conditions without serious effect such as crashing or becoming locked in a loop.

Structure: Structured programming is a methodology that forces the programmer to use certain clear, well defined methods of program design, rather than allow complete freedom to design intricate, complex programs which are prone to error and difficult to understand and debug.

Modularity: Modular programming breaks a program down into smaller, individual blocks or modules each one of which can be separately specified, written and tested. This makes the program much easier to understand and check.

Fault tolerance: Programs should be written so that if an error does occur, the program should be able to find its way out of the error condition and indicate the source of the error. In application where safety is vital, for example where a computer is used to control an industrial process, if an error does occur the program should be first detect it, send an alarm message and then move the process to known safe conditions. Fault tolerance can also be provided using program redundancy.

Languages: The reliability of software also depends on the computer language used. Assembly level programs are faster to run and require less memory than high level programs and may be preferred for real time systems. High level languages, for example, Fortran, basic, Ada, Pascal, C++, Java are processor independent and operate using advanced compilers. The newer high level languages strongly encourage structured programming.

Software checking: Modern high level language compilers have error detection capability so that errors of logic, syntax and other coding errors can be detected and corrected before an attempt is made to load the program.

The above methods and practice in each and every phase of software development cycle improves system reliability.

VI. CONCLUSION

In this paper authors are interested in software quality and reliability metrics which are two faces on the same coin for a software product measurement. Software reliability is a key concern of many users and developers of software, because reliability is defined in terms of failures, it is impossible to measure before development is complete. The approach to quantify reliability metrics is computationally intensive than simply adopting a single techniques. We explained new methods intuitively by calculating the reliability and availability of a range of practical systems given the reliability of the constituent elements and components. The authors also explained novel concepts to improve the software reliability. The ultimate outcome is a trustworthy, worthwhile, a state of art approach to reliability assessment for complex computer systems.

REFERENCES

[1] Chapin N, *A measure of software complexity* , Proceedings of the National Computer conference, pp. 728- 789, Nov 2005.

[2] Courteny R.E, *Shortgun correlations in software measures*, Software Engineering Journal, vol.8 (1), pp. 5-13, Aug 2003.

[3] Fenton N.E, *Metrics and Software structure*, Journal of Information and software technology, vol 5, pp. 20-23, 2006

[4] Haslehust M, *Manufacturing Technology*, English Universities Press, London, pp. 355-362, 2004

[5] Kolarik W. J, C*reating quality: Concepts and Systems*, McGraw-Hill, New York, pp. 879-82, 2003.

[6] National Semiconductor Corporation, *The reliability Handbook*, 3rd edition, NSC California, pp.18-33, 2000

[7] Smith D.J, *Reliability and Maintainability in perspective*, 3rd edition, Macmillan, Basingstoke, pp.9-11, Jan 2002.

[8] Taguchi G, *Experimental designs*, 3rd edition, 2 Vols,Maruzen, Tokyo, pp. 25, May 2005

[9] Thomson J.R, *Engineering Saftey assessment*, Longman Scientific and Technical Manual, Harlow, pp. 10-13, 2001

[10] Weyuker E.J, *Evaluating software complexity measures* , IEEE Transactions on Software Engineering, SE-14(9), pp. 56 -78, 2002.

AUTHORS PROFILE

L.Rakesh received his M.Tech in Software Engineering from Sri Jayachamarajendra College of Engineering, Mysore, India in 2001 as an honour student. He is a member of International Association of Computer Science and Information Technology, Singapore. He is also a member of International Association of Engineers, Hong Kong. He is pursuing his Ph.D degree from Magadh University, India. Presently he is working as a Assistant professor and Head of the Department in Computer Science & Engineering, SCT Institute of Technology, Bangalore, India. He has presented and published research papers in various National, International conferences and Journals. His research interests are Formal Methods in Software Engineering, 3G-Wireless Communication, Mobile Computing, Fuzzy Logic and Artificial agents.

Dr.Manoranjan Kumar Singh received his Ph.D degree from Magadh University, India in 1986. This author is Young Scientist awardee from Indian Science Congress Association in 1989. A life member of Indian Mathematical society, Indian Science congress and Bihar Mathematical society. He is also the member of Society of Fuzzy Mathematics and Information Science, I.S.I. Kolkata. He was awarded best lecturer award in 2005. He is currently working as a Senior Reader in post graduation Department of Mathematics, Magadh University. He is having overall teaching experience of 26 years including professional colleges. His major research Interests are in Fuzzy logic, Expert systems, Artificial Intelligence and Computer-Engineering. He has completed successfully Research Project entitled, Some contribution to the theory of Fuzzy Algebraic Structure funded by University Grants Commission, Kolkata region, India. His ongoing work includes Fuzzification of various Mathematical concepts to solve real world problems.

Dr. Gunaseelan Devaraj received his PhD degree from PSG College of Technology, Bharathiar University, Coimbatore, India in 2001. He is a member in more than 10 International and National associations in the field Computer science and Information Technology. Presently he is working as Professor and Head, Department of Information Technology, Ibri College of Technology, Sultanate of Oman since September 2007. He has more than 22 years of teaching, research and administrative experiences in Tamilnadu State Government College, Central Government University, India and Technical College in Sultanate of Oman. He has successefully completed many research projects in State Government and Central Government, India. He has presented and published research papers in various National, International conferences and Referred Journals. His area of research is Software Engineering, Algorithms in Data Mining, Bioinformatics and Computational Molecular Biology.

Preprocessing of video image with unconstrained background for Drowsy Driver Detection

M.Moorthi[1], Dr. M.Arthanari[2], M.Sivakumar[3]

[1] Assistant Professor, Kongu Arts and Science College, Erode – 638 107, Tamil Nadu, India

[2] Prof. & Head, Tejaa Sakthi Institute of Technology for Women, Coimbatore – 641 659, Tamil Nadu, India

[3] Doctoral Research Scholar, Anna University, Coimbatore, Tamil Nadu, India

Email: moorthi_bm_ka@yahoo.com, arthanarimsvs@gmail.com, , Email: sivala@gmail.com

Abstract

The face recognition includes enhancement and segmentation of face image, detection of face boundary and facial features, matching of extracted features, and finally recognition of the face. Though a number of algorithms are devised for face recognition, the technology is not matured enough to recognize the face of a person since the algorithm deal with significant amount of illumination variation in image. We propose a new image preprocessing algorithm that compensates for the problem. The proposed algorithm enhances the contrast of images by transforming the values in an intensity image, so that the histogram of the output image is approximately uniformly distributed on pixel. Our algorithm does not require any training steps or reflective surface models for illumination compensation. We apply the algorithm to face images prior to recognition. Simulation is done using seventy five web camera images using Mat lab 7.0.

Keywords: Facial recognition, Facial features extraction, Eye detection

1. Introduction

The preprocessing of real image is a crucial aspect in many useful applications like video coding of faces for video phony, animation of synthetic faces, driver behaviors analysis, word visual recognition, expression and emotion analysis, tracing and recognition of faces. The detection of facial features has been approached by many researchers and a variety of methods exist. Nevertheless, due to the complexity of the problem and illumination changes, robustness and preprocessing steps of these approaches are still a problem. Most commonly, natural face feature templates taken from real person are used for a template matching algorithm [1],[2]. These templates have to satisfy a set of requirements like orientation, size and illumination. Therefore preprocessing step is necessary for at least aligning and size changes. A wavelets based approach is described in [3]. Face images and face features from a database have to be aligned in orientation and size in preprocessing step. Both previous described methods are limited by the used template and face database.

In this paper we propose a novel low cost method designed for preprocessing. The preprocessing has three steps. In first steps modified histogram equalization is used to enhance the brightness and contrast of the images. In steps two, median filter is used to remove salt and pepper noise. Third, Binary image are obtained through the thresholding.

This paper is organized as follows. Literature surveys are given in section 2. In section 3 we will devote ourselves to discussing the preprocessing method in detail. Experimental results are reported in section 4. Conclusions will be drawn in section 5.

2. Literature Survey

Besides pose variation, illumination is the most significant factor affecting the appearance of faces. Ambient lighting changes greatly within and between days and among indoor and outdoor environments. Due to the 3D shape of the face, a direct lighting source can cast strong shadows that accentuate or diminish certain facial features. Evaluations of face recognition algorithms consistently show that state-of-the-art systems can not deal with large differences in illumination conditions between gallery and probe images [1-3].

The face detection algorithms are based on either gray level template matching or computation of geometric relationships among facial features. In recent years many appearance-based algorithms have been proposed to deal with the problem [4-7]. Belhumeur showed [5] that the set of images of an object in fixed pose but under varying illumination forms a convex cone in the space of images. The illumination cones of human faces can be approximated well by low-dimensional linear subspaces [8]. The linear subspaces are typically estimated from training data, requiring multiple images of the object under different illumination conditions. Alternatively, model-based approaches have been proposed to address the problem. Blanz et al. [9] fit a previously constructed morphable 3D model to single images. The algorithm works well across pose and illumination, however, the computational expense is very high.

In general, an image $I(x; y)$ is regarded as product $I(x; y) = R(x; y)L(x; y)$ where $R(x; y)$ is the reflectance and $L(x; y)$ is the illuminance at each point $(x; y)$ [10]. Computing the reflectance and the illuminence fields from real images is, in general, an ill-posed problem. Therefore, various assumptions and simplifications about L, or R, or both are proposed in order to attempt to solve the problem. A common assumption is that L varies slowly while R can change abruptly. For example, Homomorphic filtering [11] uses this assumption to extract R by high-pass filtering the logarithm of the image. In this paper,

enhanced method of histogram equalization is used to preprocess the image.

3. Preprocessing

In order to obtain appropriately-segmented binary images, an image preprocessing is applied. To compensate for illumination variations and to obtain more image details, modified histogram equalization is used to enhance the brightness and the contrast of the images. Then a median filter is used to remove the noise. Binary images are obtained through thresholding. The preprocessing steps are shown in Fig 1.Let we see the steps of preprocessing one by one.

3.1. Capturing image

The required images are taken from the video image using web camera.

3.2 Enhancing the image

Histogram equalization is a method in image processing of contrast adjustment using the image's histogram. This method usually increases the global contrast of many images, especially when the usable data of the image is represented by close contrast values. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to gain a higher contrast without affecting the global contrast. Histogram equalization accomplishes this by effectively spreading out the most frequent intensity values.

The method is useful in images with backgrounds and foregrounds that are both bright or both dark. A key advantage of the method is that it is a fairly straightforward technique and an invertible operator. So in theory, if the histogram equalization function is known, then the original histogram can be recovered. The calculation is not computationally intensive.

Histogram equalization often produces unrealistic effects in photographs; however it is very useful for scientific images like thermal, satellite or x-ray images, often the same class of images that user would apply false-color to

```
┌─────────────────┐
│     Input       │
│     Video       │
│                 │
└────────┬────────┘
         │
         ▼
┌─────────────────┐
│   Enhancing     │
│    video        │
│                 │
└────────┬────────┘
         │
         ▼
┌─────────────────┐
│    Median       │
│   filtering     │
└────────┬────────┘
         │
         ▼
┌─────────────────┐
│  Thresholding   │
│                 │
└────────┬────────┘
         │
         ▼
┌─────────────────┐
│    Binary       │
│    image        │
└─────────────────┘
```
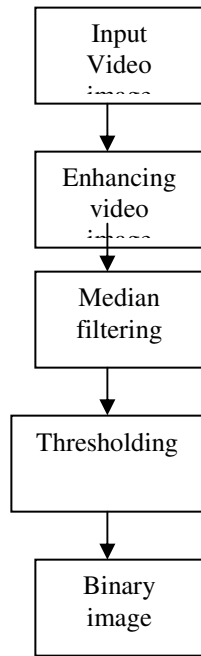
Fig. 1: Preprocessing steps.

. Also histogram equalization can produce undesirable effects (like visible image gradient) when applied to images with low color depth. For example if applied to 8-bit image displayed with 8-bit gray-scale palette it will further reduce color depth (number of unique shades of gray) of the image. Histogram equalization will work the best when applied to images with much higher color depth than palette size, like continuous data or 16-bit gray-scale images.

There are two ways to think about and implement histogram equalization, either as image change or as palette change. The operation can be expressed as *P(M(I))* where *I* is the original image, *M* is histogram equalization mapping operation and *P* is a palette. If we define new palette as *P'=P(M)* and leave image I unchanged then histogram equalization is implemented as palette change. On the other hand if palette P remains unchanged and image is modified to *I'=M(I)* then the implementation is by image change. In most cases palette change is better as it preserves the original data.

Generalizations of this method use multiple histograms to emphasize local contrast, rather than overall

contrast. Examples of such methods include adaptive histogram equalization and contrast limiting adaptive histogram equalization.

Histogram equalization also seems to be used in biological neural networks so as to maximize the output firing rate of the neuron as a function of the input statistics. This has been proved in particular in the fly retina. Histogram equalization is a specific case of the more general class of histogram remapping methods. These methods seek to adjust the image to make it easier to analyze or improve visual quality

3.3. Proposed Modification

While the results of a standard histogram equalization filtering over the whole image just described give promising results, we wanted to see if the results could be further improved. Many well-known enhancement algorithms such as histogram equalization and homomorphic filtering are global in nature and are intended to enhance an image and deal with it as a whole. We tried to split the original image in sub-images and filter each sub-image individually. First we decided to try and split the image into two halves vertically (thus obtaining two sub-images of the original image) and then apply the filter to each half individually. Second idea was to split the image horizontally and again apply the filter to each half individually. Encouraged by the good results obtained with both these methods (see Section 4 for details) we further tried to combine the filtering results into a joint representation. Let $I_{HEV}(x,y)$ be the image split vertically and each half filtered with histogram equalization filter individually, let $I_{HEH}(x,y)$ be the same for horizontally split images and let $I_{HEMOD}(x,y)$ be our proposed modification:

$$I_{HEMOD}(x, y) = 0.5[\, I_{HEV}(x, y) + .70\, I_{HEH}(x,y)]$$

Since *IHEV* scored higher results than *IHEH* in our tests we decided to keep the whole *IHEV* and multiply *IHEH* with a constant of 0.70 (chosen based on experimental results),

147

to lower its influence on the final representation. This combination produced highest results in our experiments and was kept as a final representation. We will show in the following section that our method yields superior results, and therefore justifies further research of the histogram equalization filtering variations as a means of simple yet efficient image preprocessing.

As shown in Fig. 4 (a), the input image has low contrast due to illumination; segmentation results, therefore, are unlikely to be good. Fig. 4 (b) demonstrates the image enhanced by modified histogram equalization the contrast is improved, and the details in the face region are enhanced which are discussed in detail in the following section.

3.4 Median filtering

The intensity in the eye region and other facial features is dark in a grey-level facial image. The image has been enhanced through modified histogram equalization. In image processing it is usually necessary to perform a high degree of noise reduction in an image before performing higher-level processing steps, such as edge detection. The median filter is a non-linear digital filtering technique, often used to remove noise from images or other signals.

Median filtering is a common step in image processing. It is particularly useful to reduce speckle noise and salt and pepper noise. Its edge-preserving nature makes it useful in cases where edge blurring is undesirable.

3.5 Thresholding

After median filtering threshold is set to 128, so that only dark pixels remain, including eye pair structure. Then, a binary image is obtained, which obviously contains the facial structure. Taking into account that the nonface area can influence the speed and the results of template matching, the oversize black area, which is useless in the binary image, is eliminated by the conventional connected components labeling process.

3.6 Binary image

Binary images are obtained through the thresholding. Then the final feature image is obtained, as shown in Fig. 4 (c).

4. Experimental Results

The proposed method was tested on the real video images. The video image of [480 x 640 pixels] of 75 different test persons and has been recorded during several sessions at different places. This set features a larger variety of illumination, background and face size. It stresses real world constraints. So it is believed to be more difficult than other datasets containing images with uniform illumination and background. The facial image can be preprocessed successfully in most cases, no matter whether face patterns are in different scale, expression, and illumination conditions. Typical results of preprocessing with the proposed approach are shown in Fig.4. The input images vary greatly in background, scale, expression and illumination, the images also including partial face occlusions and glasses wearing.

4.1 Method Tested

No enhancement (NE). For this test we only geometrically normalized the images (actually, images were geometrically normalized in all subsequent tests as well). No filtering or histogram equalization is used.

Standard histogram equalization (HE): Images were geometrically normalized and a standard histogram equalization (HE) technique was employed. HE enhances the contrast of images by transforming the values in an intensity image, so that the histogram of the output image is approximately uniformly distributed on pixel intensities of 0 to 255.

HE vertical (HEV): Histogram equalization filtering of two sub images are obtained by vertically dividing the input image

into two halves prior to filtering and then filtering each of them. The resulting image is obtained by concatenating the two filtered halves.

HE horizontal (HEH): The same procedure as in HEV is used with the exception of an image being horizontally divided.

HE modified (HEMOD): Method proposed in Section 3.3, consisting in combining results from HEV and HEH

Table 1, Results of applying all the techniques on video images The numbers in the table represent rank 1 recognition rate (RR) in percentages of correctly recognized images over the whole probe set.

Method	NE	HE	HEV	HEH	HEMOD
RR %	4.15	48.20	60	58.30	60.20

Table 1: Recognition rate in percentages

4.2 Results

The fig. 2 shows the proposed preprocessing method gives better results for finding the correct eye than other method since the recognition rate of the eye here is 60.2%.
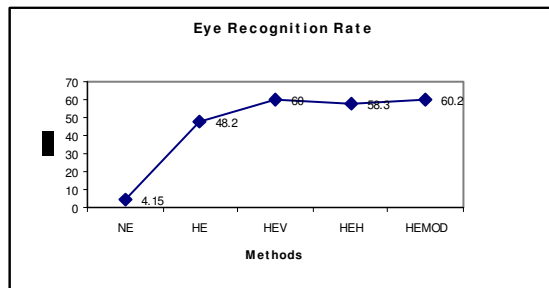
Fig. 2 Comparison of various methods
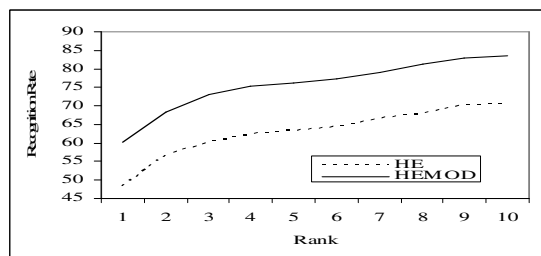
Fig. 3 Cumulative Match score curves for HE and proposed method

The fig 4 shows the implementation results of image preprocessing. The fig 4.a and fig 4.b are the results of enhanced image and binary image. By looking at the extremely low recognition rate on NE images just 4.15%, the proposed method is better. In our experiment the standard preprocessing HE which yielded only 48.20%. HEV and HEMOD give significant improvement with 60% and 60.20% respectively. Therefore, we can see clearly that our proposed method is superior to all other methods and recognition rate is 12% is higher than the standard HE. The superiority of the proposed method is further confirmed in Fig. 3 where the cumulative match score curve for the standard method and proposed method could seen.

(a)   (b)   (c)

(a)   (b)   (c)

Fig. 4: An example of preprocessing (a) Original Image (b) Enhanced image (c) Binary image

5. Conclusions

We introduced a simple image-preprocessing algorithm for compensation of illumination variations in images. The algorithm enhances the contrast of images by transforming the values in an intensity image so that the histogram of the output image is approximately uniformly distributed on pixel intensities. The algorithm delivers large performance improvements for standard face recognition algorithms. Experiments demonstrated the robustness of the method with several images captured from web camera.

References

1. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for Face recognition algorithms. IEEE PAMI 22 (2000) 1090-1104

2. Blackburn, D., Bone, M., Philips, P.: Facial recognition vendor test 2000: evaluation report (2000)

3. Gross, R., Shi, J., Cohn, J.: Quo vadis face recognition? In: Third Workshop on Empirical Evaluation Methods in Computer Vision. (2001)

4. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigen faces vs. Fisherfaces: Recognition using class specific linear projection. IEEE PAMI 19 (1997) 711-720

5. Belhumeur, P., Kriegman, D.: What is the set of images of an object under all possible lighting conditions. Int. J. of Computer Vision 28 (1998) 245-260

6. Georghiades, A., Kriegman, D., Belhumeur, P.: From few to many: Generative models for recognition under variable pose and illumination. IEEE PAMI (2001)

7. Riklin-Raviv, T., Shashua, A.: The Quotient image: class-based re-rendering and recognition with varying illumination conditions. In: IEEE PAMI. (2001)

8. Georghiades, A., Kriegman, D., Belhumeur, P.: Illumination cones for recognition under variable lighting: Faces. In: Proc. IEEE Conf. on CVPR. (1998)

9. Blanz, V., Romdhani, S., Vetter, T.: Face identification across different poses and illumination with a 3D morphable model. In: IEEE Conf. on Automatic Face and Gesture Recognition. (2002)

10. Horn, B.: Robot Vision. MIT Press (1986)

11. Stockam, T.: Image processing in the context of a visual model. Proceedings of the IEEE 60 (1972) 828-842

12 A. V. Oppenheim, R. W. Schafer, and T. G. S. Jr, "Nonlinear filtering of multiplied and convolved signals" IEEE Proc., vol. 56, no. 8, pp. 1264–1291, 1968.

Mr. M. Moorthi received MCA degree from Bharathiar University, Coimbatore, and M.Phil., Degree from Manonmaniam Sundaranar University and doing PhD Degree in R&D centre, Bharathiar University, Coimbatore, TN, India . He is currently the Lecturer (SG) in Kongu Arts and Science College, Erode, TN, India. He has 11 years of teaching and 6 years of research experience. He has guided nine M.Phil students in the area of Computer Science. He has presented papers in National and International Conference and has published an article in National Journal. He is a member of ISCA and working as Associate Editor in Canadian Research & Development Center of Science and Cultures – Advances in Natural Science and Management Engineering – ISSN 1913-0341.His interests and expertise are in the area of Image Processing, Data Mining, Multimedia, Computer Graphics and Networks.

E-mail ID: moorthi_bm_ka@yahoo.com.

Dr Arthanari holds a Ph.D in Mathematics from Madras University as well as Masters degree in Computer Science from BITS, Pilani. He holds patent issued by the Govt. of India for the invention in the field of Computer Science. He has directed teams of Ph.D researchers and industry experts for developing patentable products. He teaches strategy, project management, creative problem solving, innovation and integrated new product development for last 35 years.

Mr. Sivakumar M (sivala@gmail.com) has 10+ years of experience in the software industry including Oracle Corporation. He received his Bachelor degree in Physics and Masters in Computer Applications from the Bharathiar University, India. He holds patent for the invention in embedded technology. He is technically certified by various professional bodies like ITIL, IBM Rational Clearcase Administrator, OCP - Oracle Certified Professional 10g and ISTQB.

Ultra Fast Computing Using Photonic Crystal Based Logic Gates

X.Susan Christina
Dept. of ECE
Mookambigai College of Engg.
Trichy- 622 502, India.
fab_jesu@yahoo.co.in

A.P.Kapilan
Dept. of ECE
Chettinad College of Engg & Tech
Karur,. 639114. India.
apkabilan@yahoo.co.in

P. Elizabeth Caroline
Dept. of ECE
JJ College of Engg &Tech,
Trichy –620 009,India.
becaroline05@yahoo.com

Abstract— **A Novel design of all-optical fundamental NAND and XNOR logic gates based on two dimensional photonic crystals has been presented in this paper. In a photonic crystal self collimated beams are partially transmitted and partially reflected with a phase lag at line defect in Γ-X direction. By employing a appropriate phase shifter, the reflected and transmitted input beams are interfered constructively or destructively to obtain the required logic outputs**. The operation of the logic gates is simulated using two dimensional Finite Difference Time Domain (FDTD) method.

Keywords-optical computing; logic gated; photonic crystal; self collimated beam; FDTD

I. INTRODUCTION

The demand for bandwidth in worldwide networks continues to increase due to growing internet usage and high bandwidth applications. Optical computing is one of promising technique to meet all the necessary requirements such as high speed, high speed, supporting high data rate and ultra fast performance [1,2]. All optical logic gates are the key element in next generation optical computing and in networking to perform optical signal processing such as binary addition, header reorganization, parity checking, optical bit pattern recognition addressing, demultiplexing, regenerating and switching. In order to realize the gates, various configurations have been reported that utilize the nonlinear properties of the optics. All-optical gates reported in the literature [3-8] could be achieved with a semiconductor laser amplifier loop mirror (SLALOM), a semiconductor optical amplifier- (SOA-) based Mach-Zehnder interferometer (SOA-MZI), a SOA based ultra fast nonlinear interferometer (UNI), cross-polarization modulation, and four-wave mixing (FWM) in SOAs, SOA with Optical filter, Periodically Poled Lithium Noibate (PPLN) waveguide . These schemes suffered from certain fundamental limitations such as spontaneous emission noise, power consumption and size.

In recent years, optical waveguide element employing photonic crystals have been received lot of attention because of their dimension, low loss structure of less than 2 dB/cm [9] and high speed with data rate of 120 GB/s [10]. Normally

Photonic Crystals (PC) are produced by artificially imparting periodic change of the refractive index of a structure which has a band gap that prevents propagation of certain frequency range of light. But the propagation of light inside the PC can be controlled by different propagation mechanisms such as negative refraction, super prism and self collimated beam propagation. When non linear effect is applied to the photonic crystal it requires high intensity incident light for its propagation and the balance between diffraction and focusing easily collapses due to the absorption. In self-collimating effect, the collimated light beam insensitive to the divergence of the incident beam without applying a nonlinear effect [11]. In this paper we propose NAND and XNOR gates realization.

The paper is organized as follows, In Section II, photonic crystal theory is described. In Section III, structural and numerical analysis is explained. Section IV presents the proposed scheme of logic gates. Results and related discussions are presented in section V. Finally, conclusions are summarized in section VI.

II. PHOTONIC CRYSTAL THEORY

Photonic crystals (PC) are composed of periodic dielectric materials. In PC, for some frequency ranges the light waves are not propagating through the structure such frequency range is called forbidden gap photons. The doping of impurity or creating defects will allow a perfect control of light propagation and radiation. Introducing line defects in PC results in a photonic crystal waveguide. Line defects can be formed in photonic crystal either by reducing the radii of PC rods or by eliminating them partially. When the self-collimated beam is incident at the line defect the beam is splitted [12, 13]. It is evident that there is a phase difference between the transmitted and the reflected beams. If the rod radii of the line defect are smaller than that of the host rod radii, the reflected wave lag the transmitting wave by π/2 else the phase difference is - π/2 [14]. If another self- collimated beam with appropriate phase is launched, the reflected and transmitted beams may interfere constructively or destructively. This phenomenon is used to realize logic gates functions.

III. STRUCTURAL AND NUMERICAL ANALYSIS

To realize the operation of the all optical logic gates, a 2D square lattice PC composed of silicon dielectric rods in air is considered. The size of the PC is 6.4 x 6.4 μm. The refractive index of the silicon rod is 3.5. The radius and the dielectric constant of rods are 'r' =0.35a and 'ε' = 12.0 respectively [12], where 'a' is the lattice constant and its value is 0.365 μm. The line defect is formed by reducing the silicon rod radii = 0.274a of 15 rods aligned in the Γ-X direction. Self collimation phenomena occurs when lights of frequencies around f = 0.194 c/a [12] where 'c' is the speed of light in free space propagate along the direction of Γ-M. Fig 1 shows the schematic diagram of the Photonic crystal. In this structure there are four faces, two of them are consider as input and remaining two are as output.
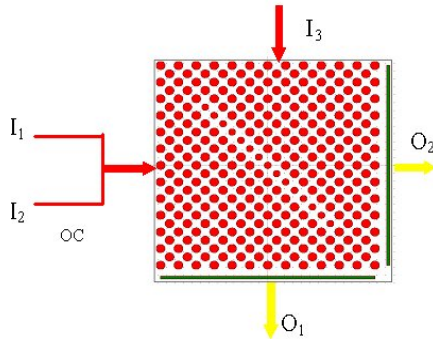
Figure 1. Schematic diagram of photonic crystal.

To analysis photonic crystal, FDTD with perfectly matched layer boundary condition method is used in this paper. It solves Maxwell's equations by first discretizing the equations via central differences in time and space and then numerically solving these equations. Since the whole calculation region is divided into very small uniform cells, the accuracy of this technique can be improved. Photonic wave guides are very small due to the frequency of light. It is both expensive and complicated to construct these. Therefore FDTD simulation is a great interest to analysis. Fig. 2 depicts the band diagram of the PC using FDTD simulation.
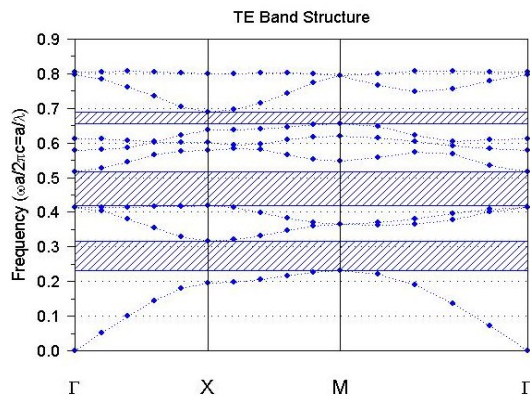
Figure 2. Band diagram of photonic crystal.

IV. PROPOSED SCHEME OF LOGIC GATES

A. Schematic of XNOR Gate

Logic gates function can be realized by introducing a certain phase difference between the input beams. To realize XNOR gate along with two input beams, the third reference input beam is also incident on the PC. The inputs I_1 and I_2 are launched at the input face 1 and the third reference beam is applied to the input face 2. The optical phase shifter is connected at the reference input to obtain appropriate phase shift. The phase difference between the inputs I_1 and I_2 are zero i.e. φ_1- φ_2= 0 and the phase difference between inputs and the reference input φ_1- φ_3 is set as $\pi/2$. The XNOR output is taken from the output face 2.
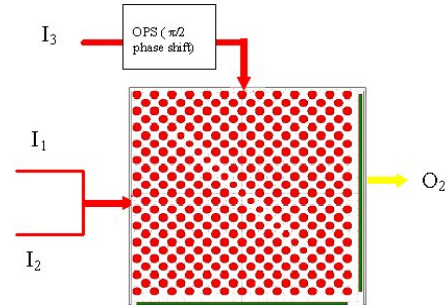
Figure 3. Schematic of XNOR gate.

B. Schematic of NAND Gate

The gate NAND can be realized by applying the input beams I_1 and I_2 on input face 1 and the reference beam is launched at input face 2. The inputs powers consider in this case are half of the reference input power. The phase difference between the inputs I_1 and I_2 is zero i.e. φ_1- φ_2= 0 and the phase difference between inputs and the reference input φ_1- φ_3 set as $\pi/2$. The NAND output is taken from output face 2.

Figure 4. Schematic of NAND gate.

V. RESULTS AND DISCUSSIONS

In the XNOR gate when two input beams and the reference input with the phase difference $\pi/2$ are introduced, the output light will be at the face O_2. If only one input with reference input is applied, there is no output at the face O_2. Table 1 gives

the functions of the XNOR gate and Fig. 5 shows the field distributions of TE mode for various input combinations.

various input combinations are shown in Fig. 6 and Table 2 gives the functions of the NAND gate.

TABLE I. FUNCTIONS OF XNOR GATE

Signal Descriptions	XNOR for $\varphi_1 - \varphi_2 = 0$ & $\varphi_1 - \varphi_3 = \pi/2$ and the input powers $I_1 = I_2 = I_3$			
Input signal (I_1)	0	0	1	1
Input signal (I_2)	0	1	0	1
Control signal (I_3)	1	1	1	1
Output O_2	1	0	0	1

TABLE II. FUNCTIONS OF NAND GATE

Signal Descriptions	NAND gate for $\varphi_1 - \varphi_2 = 0$ & $\varphi_1 - \varphi_3 = \pi/2$ and the input powers $I_1/2 = I_2/2 = I_3$			
Input signal (I_1)	0	0	1	1
Input signal (I_2)	0	1	0	1
Control signal (I_3)	1	1	1	1
Output O_2	1	1	1	0

Figure 5a) Simulated field distribution when both inputs are high.

Figure 6a) Simulated field distribution when both inputs are high.

Figure 5b) Simulated field distribution when one of the input is high.
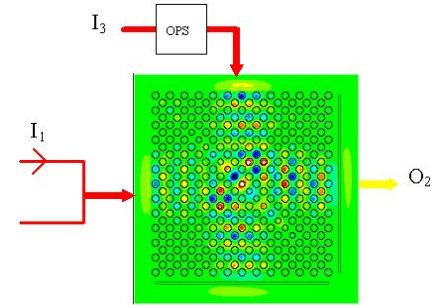
Figure 6b) Simulated field distribution when one of the input is high.

VI. CONCLUSION

Figure 5c) Simulated field distribution when both inputs are low.

In the NAND gate when two input beams whose powers are half of the reference input power is introduced, no output signal is from the face O_2. If only one input with reference input or only reference input is applied, there is an output signal in the face O_2. The TE mode field distributions for

The design of novel all-optical logic gates consisting of phase shifter and photonic crystal with a line defect in the Γ-X direction is proposed. The self-collimated optical beams are applied at a line defect of the photonic crystal that are partially transmitted and reflected with a phase lag. If the intensities of the input beams are chosen in appropriate proportions and opposite phase difference between the input signals, the overlapping of transmitted and reflected beams interfere either constructively or destructively giving a logic output. Based on these phenomena the XNOR and NAND gates functions are realized. The steady state field distributions at different input states are obtained by FDTD simulation. The results indicate that photonic crystals are potential candidature for optical digital integrated circuits which are used for optical computing.

REFERENCES

[1] J.M. Martinez,J.Herrera, F.Ramos and J. Marti, " All-Optical Address Recognition Scheme for Label-Swapping Networks," IEEE Photonics Technology Letters, vol. 18, no.1, pp. 151-153,2006.

[2] T.Fujiswa and M. Koshiba, "Finite-Element Modelling of non linear Mach-Zehnder interferometers based on photonic crystal wave guides for All-Optical signal Processing," Journal of Lightwave Technology, vol. 24, no. 1pp. 617-623, 2006.

[3] A. J. Poustie, K. J. Blow, A. E. Kelly, and R. J. Manning, "All optical full adder with Bit-differential delay," Opt.Commun., vol.156, pp.22-26, 1998.

[4] A. J. Poustie, K. J. Blow, A. E. Kelly, and R. J. Manning, "All optical parity checker with bit-differential delay,"Opt. Commun., vol.162, pp.37-43, 1999.

[5] H. Avramopoulos, "Optical TDM devices and their applications," Optical Fiber Communication, vol. 54, 2001.

[6] H. Soto, C. A. Díaz, J. Topomondzo, D. Erasme, L. Schares and G. Guekos, "All-Optical AND gate implementation using cross polarization modulation in a semiconductor optical amplifier," IEEE Photon. Technol. Lett.,vol. 14, pp.498-500, 2002.

[7] D. Nesset, M. C. Tatham, and D. Cotter, "All Optical gate operating 10 Gbits/s signals at the same wavelength using four-wave mixing in a semiconductor laser amplifier", Electronics Letters, vol.31, no. 31, pp.896-897, 1995.

[8] J.Wang, J.Sun and Q.Sun "PPLN based Flexible Optical Logic AND gate," IEEE Photon Tech Lett. vol. 20, pp. 211-213, 2008.

[9] E. Kuromochi, M. Notomi, S.Hughes et al, "Disorder-induced scattering loss of the line defect waveguides in photonic crystal slabs," Phys. Rev.B, vol.72, 161318(R), 2005.

[10] Parisa Andalip and Nosratollah Granpayeh, "All-Optical ultra-compact photonic crystal AND based on nonlinear ring resonators," Journal of Opticla Society of America B, vol.26, no.1,pp. 10-16, 2009.

[11] H.Kosaka, T.Kawashima, A.Tomita, M.Notomi, T.Tamamura, T.sato and S.Kawakami, "Self-collimatin phenomena in photonic crystal', Appl.Phys.Lett.,vol.74, pp.1212-1214, 1999.

[12] X. Yu and S. Fan, "Bends and splitters for self-collimated beams in photonic crystals," Appl. Phys. Lett., vol. 83, 3pp. 251-3253, 2003.

[13] S.G.Lee, S.S.Oh, J.E.Kim, H.Y.Park and C.S.Kee, "Line-defect-induced bending and splitting of self-collimated beams in two-dimensional photonic crystal," Appl. Phys. Lett., vol. 87, 181106-3, 2005.

[14] D. D. Zhao, J. Zhang, P. Yao, X. Jiang, and X. Chen," Photonic crystal Mach - Zehnder interferometer based on self-collimation," Appl. Phys. Lett., vol.90, 231114-1, 2007.

AUTHORS PROFILE

X. Susan Christina pursuing her Doctorate in Optical Computing at Anna University, Trichy, India. She has 13 years of teaching experience. She has been working as a Professor and Head of the Department of Electronics and Communication Engineering at Anna University, Trichy. Her research interests in the area of digital signal processing, optical signal processing and optical computing. She has published fifteen papers in National, International conferences proceedings and Journals. She is a life member of ISTE.

Dr. A.P. Kabilan has a Doctorate in Microwave Engineering from Patrice Lumumba University, Moscow, Russia and has 25 years of experience in teaching and research. He is a dynamic academician with 26 international publications and eight national ones. He worked as a Professor and Head of the Department of Electronics and Communication Engineering at Anna University, Coimbatore for the past ten years. At present, he is the Principal of Chettinadu College of Engineering and Technology, Karur, and Tamil Nadu India. He is an active member of IEEE and a life member of ISTE. He has a strong background in microwave, optical and antenna engineering.

P. Elizabeth Caroline is pursuing her Doctorate in Optical Signal Processing at Anna University, India. She has 16 years of teaching experience. She has been working as a Professor and Head of the Department of Electronics and Communication Engineering at Anna University, Trichy. She has a strong background in signal processing and optical computing. She has presented papers in international IEEE conferences and national conferences. She is an active member of IEEE and a life member of ISTE.

Markov Chain Simulation of HIV/AIDS Movement Pattern

Ruth Stephen Bature
Department of Computer/Mathematical Science
School of Science Technology
Federal College of Chemical and Leather Technology,
Zaria, Nigeria.
rsbature@yahoo.com

Obiniyi, A. A.
Department of Mathematics,
Ahmadu Bello University,
Zaria, Nigeria
aaobiniyi@yahoo.com

Ezugwu El-Shamir Absalom
Department of Mathematics,
Ahmadu Bello University,
Zaria, Nigeria
Code_abs@yahoo.com,

Sule, O. O.
Department of Computer/Mathematical Science
School of Science Technology
Federal College of Chemical and Leather Technology,
Zaria, Nigeria
bumsia@yahoo.com

Abstract

The objective of this research work is to simulate the spread of HIV/AIDS from one generation to another or from one person to another, with a view of contributing to the control of the disease. This will be accomplished using Markov Chain method with a computer program written in Java to simulate the process. This paper is also concerned with the movement pattern of HIV/AIDS from one generation to another generation over a period of 20 years. This can help professional take the probability measures of HIV/AIDS over a given period of time, within a specific area or location.

Keywords: HIV/AIDS, Markov Chain, Transition Matrix, Probability Matrix

1.1   INTRODUCTION

AIDS is a term with an official definition used for epidemiological surveillance. This means that systematic reporting of AIDS cases is useful in helping to monitor the HIV pandemic and to plan public health responses. The term AIDS is not useful for the clinical care of individual patients. In managing patients with HIV-related diseases, the aim is to identify and treat whichever HIV-related diseases are present.

There are many cases in which we would like to represent the statistical distribution of these epidemiological occurrences in a state or form that will enable us analyze the trends in their behavior by means of mathematical variables to enable us predict their future behavior. Markov Chain Models are well suited to this type of task. In this research work HIV/AIDS was analyzed using the Markov Chain model.

A Markov chain is a special type of stochastic process, which may be described as follows:

At any given time n, when the current state X_n and all previous states X1, X2, X3…Xn-1 of the process are known, the probabilities of all future states Xj (j > n) depends only on the current state Xn and does not depend on the earlier states X1, X2, X3…Xn-1.

A process in which a system changes in random manner between different states, at regular or irregular intervals is called a stochastic process. If the set of possible outcomes at each trial is finite, the sequence of outcomes is called a finite stochastic process. In a stochastic process, the first observation X1 is called the initial state of the process; and for n = 2, 3… the observation Xn is called the state of the process at time n.

A Markov chain is a stochastic process such that for n = 1,2,3… and for any possible sequence of state;

X1, X2, X3 … Xn+1, Pr (Xn+1 = Xn+1, X1 = X1, X2 = X2… Xn =Xn) = Pr (Xn+ 1 = Xn + 1 / Xn = Xn)…………………1,0,0

Maki (1989) states that the Markov chain model and the modern theory of stochastic process was developed by an outstanding Russian Mathematician called Andreevich Markov.

HIV is said to develop into full-blown AIDS when the body immune system has been destroyed and can no longer perform its function to fight off diseases that may attack the body system. It is therefore, essential to certify that HIV/AIDS are infectious disease, since they are caused by a virus and can be transmitted from person to person. These persons can be categorized into the following:

The Susceptible People (S),

The Infective People (I) and

The Clinical AIDS Cases (A).

Fig: 1. HIV/AIDS States Diagram

All these fit the conditions required for Markov chain to be used for analysis. An S person does not carry the AIDS virus but can contact it through sexual contacts with I person/people, or by sharing of needles, in Intravenous drug (IV) used, or by blood transfusion of contaminated blood; there is a chance that he/she could develop AIDS symptoms to become an AIDS case.

An AIDS, case (A) person is a person who has developed AIDS symptom or has cell count in the blood falling below200/mm^3 which result to death.

In the study of the HIV/AIDS epidemiology in terms of the mode of transmission of the epidemic, there exist three types of people regarding HIV epidemic in a given population. There are: the S (susceptible) people, the I (infective) people and the A (clinical AIDS cases) people. An S person does not carry the HIV or AIDS virus but can contact it through sexual contact with I people or AIDS cases or by sharing needles or other infected materials in IV drug use or through vertical transmission from an HIV – infected mother to her child. An I person carries the AIDS virus and can transmit the virus to S people through sexual contact or sharing contaminated needles with I people; there is a chance that he/she will develop HIV/AIDS symptoms to become an AIDS case. An AIDS case (an A person) is a person who has developed AIDS symptoms or who has CD4$^+$T cells counts in the blood falling below 200/mm^3.

2.1 Literature Review

One of the most attractive features of the natural sciences such as chemistry, Biology, Physics and Mathematics is that they can formulate principles mathematically, and from these principles, they can make predictions about the behaviour of a system.

Many related literatures were reviewed and detailed information on the study of HIV/AIDS using Markov Chain Model was outlined. The symptoms were clearly stated as follows: Coughing, Slight Fever, loss of weight, sweating while

sleeping, difficulty in breathing and feeling tired through your entire body.

The provirus called human immune deficiency virus (HIV) was isolated as the causative agent of AIDS in the United State of America by Centers for Disease Control (CDC). A blood test was then formulated to detect the virus in a person, and the virus targets in the body were established. The HIV infects a subpopulation of the thymus – derived T Lymphocytes called CD4+ Lymphocytes or T4 cells, which are helper cells. These T cells perform recognition and induction function as part of the immune response to foreign stimuli. In recognizing a foreign antigen, the CD4+ T cell plays a major role in stimulating other cells, such as the macrophages, to ingest and destroy infected cells. However, a suppressor T lymphocyte "or CD8+ T cells can also attack cells infection with a virus directly by a process called "cell – mediated cytotoxicity". Since CD4+ T cells not only have direct cytotoxic activity but also secrete factors that stimulate the proliferation of CD8+ cytotoxic T cells, and are also important in promoting cell – mediated cytotoxicity. Thus, the CD4+ T plays a central role in both humoral and cell – mediated defenses (Cooley, P. C., 1993).

When an individual is infected with HIV, the clinical response is complex, progressive and varies among individuals. Within few days of infection, an individual develops an acute mononucleosis like syndrome with fever, malaria and lymphadenopathy the swelling of the lymph glands, but symptoms abate as HIV bonds to cells with CD4+ T Receptor. HIV attacks CD4+ cells because they contain the CD4+ receptors, and kills CD4+. T – Lymphocytes level drop rapidly from a pre-infection normal level of about 1125 CD4+ T cells per ml to about 800 cells per ml, the decline proceeds at a slower pace (Cooley P. C., Hamill, P. C. and Myers L. E. 1993).

One of the perpetual dreams of mankind has always been to be able to predict the future. The regular recurrence of an epidemic and the similar shapes of consecutive epidemics of a disease have for a long time tempted people with mathematical inclination to make some kind of model (Tan W.Y., 1993 and 2000.

This chance of mathematics, in turning vague questions into precise problems, in recognizing the similar features of apparently diverse situations, in organizing information and in making predictions is vital in our social and personal lives.

It was observed that the growth and development of this model probability could be traced to two separate phenomena Viz: the needs for government to collect information on its citizenry and the development of a Mathematical Model of probability theory. Today these data are used for many purposes including apportionment and strategic decision – making (Tan, 2000).

A practical application of this decision theory approach is evident in the analysis of genetics, particularly in Sickle cell anemia, and also a practical approach to this theory is evident in the analysis of employment status particularly in Bauchi State.

2.2   MARKOV CHAIN MODEL

The term Markov chain analysis refers to a quantitative technique that involves the analysis of the current behaviour of some variables in order to predict the future behaviour of that variable.

2.2.1   Properties of Markov Chain Model

a.   An experiment has a finite number of discrete outcomes called states. The process   or experiment is always in one of these states.

The sample space;

S = {X_1, X_2, X3… Xn} remains the same for each experiment or trail,

Where X_1, X_2, X_3… Xn are states.

b.   With each additional trail, the experiment can move from its present state to any other or remains in the same state.

c.   The probability of going from one state to another in the next trail depends only on the present state or proceeding trail and not on past state and upon no other previous trails.
d.   The probability of moving from any one state to another in one step is represented in a transition matrix.

For each i and j, the probability pi, j and Xj will occur given that what occurred on the preceding trail remains constant through the sequence (stationary transition probability)

2.2.2   Probability Matrix:

State	1	2	3------------N	Sum of row
1	n_{11}	n_{12}	n_{13} ----------n_{1n}	S_1
2	n_{21}	n_{22}	n_{23} ----------n_{2n}	S_2
3	n_{31}	n_{32}	n_{33}-----------n_{3n}	S_3
N	n_{N1}	n_{N2}	----------n_{NN}	S_N

Each entry n i, j in the table refer to the number of times a transition has occurred from state i to state j. The probability transition matrix is formed by dividing each element in every row by the sum of each row.
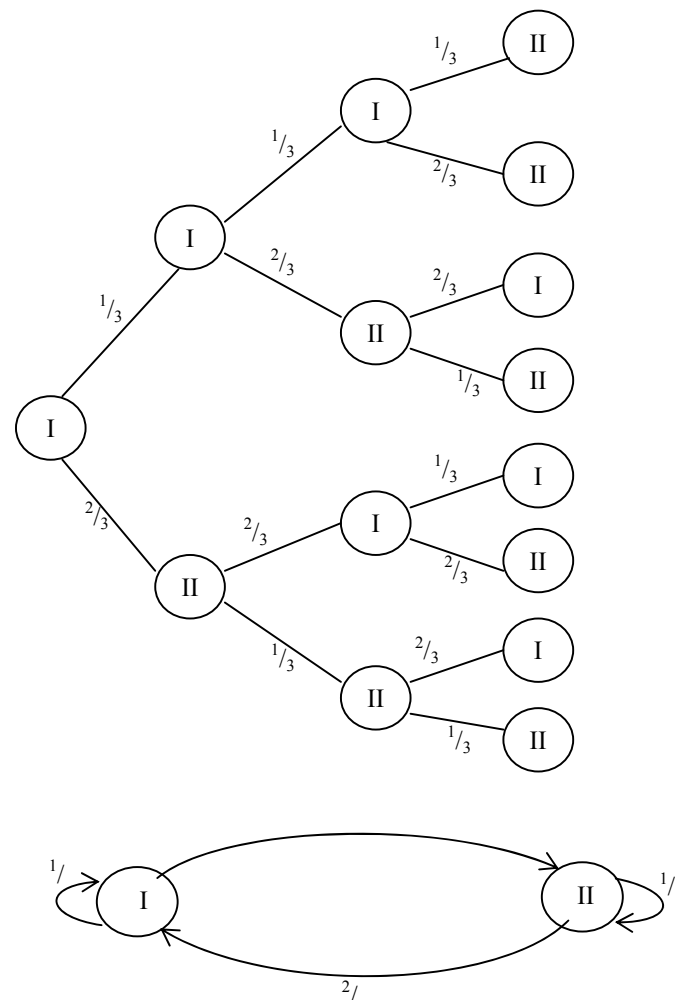
Fig.3: Transition diagram of Markov Chain

3.1   Mathematical Representation of Markov Processes

A Markov chain as earlier explained is a stochastic process such that for n = 1,2,3… and for any possible sequence of states x_1, x_2, x_3, …, $x_{n+1,}$

$Pr (X_{n+1} = x_{n+1}/X_1=x_1, X_2=x_2… X_n=x_n)=Pr (X_{n+1}= x_{n+1}/X_n=x_n)$.

From the multiplication rule of conditional probability, given as

P (E n A)=P (E) P (A/E), where E is an arbitrary event in a sample space S with P (E) > 0 and A is any event and using the fact that A n E = E n A, it follows that the probability in a Markov chain must satisfy the relation: (Norman,1961).

Pr ($X_1=x_1$, $X_2=x_2$… $X_n=x_n$) =

Pr($X_1=x_1$) Pr($X_2=x_2/X_1=x_1$) Pr($X_3=x_3/X_2=x_2$ … Pr ($X_n=x_n/X_{n-1}=x_{n-1}$)

Any vector w= (w_1, w2, w3… w_k) such that $w_i \geq 0$ for i =1, 2, 3…k and also $\sum_{i=1}^{k} w_i = 1$ is called a probability vector. The probability vector v= (v_1, v2, v3… v_k), which specifies the probabilities of the various states of a chain at the initial observation time, is called the initial probability vector for the chain.

The initial probability and the transition matrix together determine the probability that the chain will be in any particular state at any particular time. If v is the initial probability vector for a chain, then Pr ($X_1=s_i$) = V_i for i=1, 2, 3…k. If the transition matrix of the chain is the K x K matrix P having the elements $p_{i,j}$, then for j=1,2,3,…, k

$$Pr (X_2=s_j) = \sum_{i=1}^{k} Pr (X_1 = s_i \text{ and } X_2 = s_j)$$

$$= \sum_{i=1}^{k} Pr (X_1 = s_i) Pr (X_2 = s_j / X_{1 = s_i})$$

$$= \sum_{i=1}^{k} v_i P_{ij}$$

Since $\sum_{i=1}^{k} v_i P_{ij}$ is the jth component of the vector vP, this derivation shows that the probabilities for the state of the chain at the observation time are specified by the probability vector vP.

3.2   Transition Probabilities and Transition Diagraphs

Consider a finite Markov chain with K possible states S_1, S_2, S_3…,S_k and stationary transition probabilities. For i=1, 2, 3…k and j=1, 2, 3…k, and let P_{ij} denote the probability that the process will be in state s_j at a given observation time, if it is in state s_i at the preceding observation time. The transition matrix of the Markov chain is defined to be the

K x K matrix P with element p_{ij}. Thus:

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1k} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2k} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3k} \\ & & \ddots & & \\ p_{k1} & p_{k2} & & \cdots & p_{kk} \end{pmatrix}$$

Since each number P_{ij} is a probability, then $P_{ij} \geq 0$. Furthermore, $\sum_{j=1}^{k}$ for

i = 1,2,3, …., k , because if the chain is in state s_i at a given observation time, then the sum of the probabilities that it will be in each of the state's s_1,…, s_k at the next observation time must be 1.

A square matrix for which all elements are non-negative and the sum of the elements in each row is 1 is called a stochastic matrix. It is seen that the transition matrix P for any finite Markov chain with stationary transition probabilities must be a stochastic matrix. Conversely, any K x K stochastic matrix can serve as the transition matrix of a finite Markov chain with K possible states and stationary transition probabilities.

3.2.1   Transition Matrix When an I (Infected) Person Transmits the Hiv/Aids Virus to an S (Susceptible) Person through Sexual Contact

An infected person and a susceptible person here stand to include male and female individuals in a given population. The chance that a susceptible person whether a male or female can contact HIV/AIDS virus after having sexual intercourse with an infected person can be obtained, since the infected person is defined to have the HIV virus in his/her body. To obtain the possible transition composition of the stochastic matrix, consider the situation as presented in table 3.0 below.

Table 3.0 Transmission Matrix through Sexual Contact

Recipients		
Transmitters	Male (M)	Female (F)
Male (M)	P (MM)	P (MF)
Female (F)	P (FM)	P (FF)

All the infection possible in the four cells of the Table 3.0 is of heterosexual and homosexuals contacts.

The interpretation of this is that; there exist a possibility of contacting the HIV/AIDS virus when there is sexual intercourse between men (homosexuals) or intercourse between men and woman (heterosexual) as well as between woman and woman (lesbians).

This phenomenon of contacting infection will only take place if one of the sex partners is already infected with the HIV/AIDS virus as indicated in the Table 3.0.

However, to develop the transition probability, data was collected from the HIV/AIDS cases recorded at the Ahmadu Bellow University Teaching Hospital (ABUTH) Zaria for 10 years, on sexual intercourse distribution of HIV/AIDS patient. The heterosexual, homosexual and lesbians transitions for both male and female groups was simulated based on the

experimental data, and are tabulated hypothetically and illustrated as shown in table 3.1 below.

Table 3.1: Cumulative Reported Cases of HIV/AIDS through Sex

Recipients			
Transmitters	Male	Female	Total
Male	150	216	366
Female	250	26	276
Total	300	242	542

We can now develop a two-stage model from the Table 3.1, where state I denotes Male (M) transmitters and state II denotes Female (F) transmitters. To compute the probability of transition from state I to state I it is noted that 150 out of 366 or 36% of male recipients were infected through sexual intercourse between men (homosexuals). The transition from state I to state II is 250 out of 320 or 64% of female recipients, where infected through sexual intercourse between men and women (heterosexuals). Similarly, the transition from state II to state I is 275 out of 275 or 100%, meaning that, out of all the infected females who had sex with men, 100% of the men were infected. For the mode of transmission of the HIV/AIDS virus through sexual contact or intercourse, the transition from state II to state II was found to be zero (0).

Our probability transition matrix is then given by:

$$P = \begin{bmatrix} 0.36 & 0.64 \\ 1.0 & 0.0 \end{bmatrix}$$

The diagram of the transition matrix is as shown fig. 3.0:
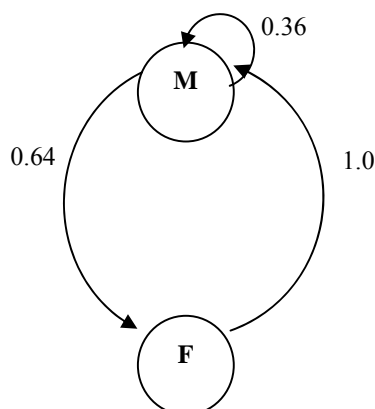
Fig. 3. 1 *Transition diagram for contacting HIV/AIDS virus through sexual contact*

3. 2.2. Transition Matrix for HIV/AIDS through Blood Transfusion

Similarly, following the same pattern of the previous section 3.2.1, the chances that a susceptible person can contact the HIV/AIDS virus after being transfused infected blood from an infected person can be obtained. To obtain the possible transition composition of the stochastic matrix, we consider the situation in Table 3.2

Table 3.2 HIV/AIDS transmission through blood transfusion for different blood groups

Recipients				
Transmitters	A	B	AB	O
A	P (A, A)	P (A, B)	P (A, AB)	P (A, O)
B	P (B, A)	P (B, B)	P (B, AB)	P (B, O)
AB	P (AB, A)	P (AB, B)	P (AB, AB)	P (AB, O)
O	P (O, A)	P (O, B)	P (O, AB)	P (O, O)

Here, blood groups A, B, AB and O represent four different blood groups for infected persons and susceptible persons respectively. The outcome in Table 3.2 indicates that there exist possibilities of susceptible persons being infected whenever there is transfusion of infected blood to them, but this will only occur when corresponding blood groups are transfused with the corresponding groups, except for blood group O, which is a universal donor can be transfused to other groups but other groups cannot be transfused to blood group O except O.

However, to develop the transition probabilities for Table 3.2 data was collected from the HIV/AIDS recoded cases at Ahmadu Bello University Teaching Hospital (ABUTH) for 20 years (1993 - 2009) on a case history of HIV/AIDS patients infected through blood transfusion based on reported cases at the hospital. Hence, the number of transfusions for the various blood groups that were infected were summed up to obtain the sum totals for ten years as obtained from the given data is tabulated and illustrated in Table 3.3.

Table 3.3 Cumulative reported cases of HIV/AIDS through blood transfusion for different blood groups.

Recipients					
Transmitters	A	B	AB	O	Total
A	42	0	0	0	42
B	0	60	0	0	60
AB	0	0	95	0	95
O	27	36	32	46	141
Total	69	96	127	46	338
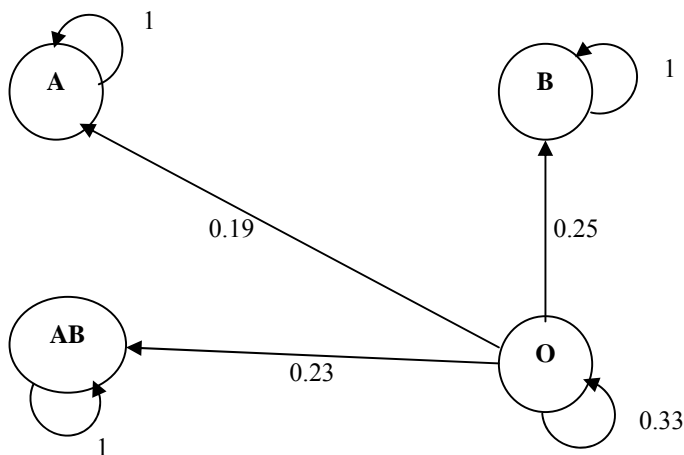
Fig 3.2: Transition diagram for contacting HIV/AIDS through blood transfusion.

3.2.3 Transition Matrix for Hiv/Aids through Vertical Transmission from a Hiv Infected Mother to her Child

Vertical transmissions of HIV infection from Mother to child occur either in utero (before birth), intrapartum (during birth) or post natal (after birth). Fetuses aborted during the first and second trimesters of pregnancy in some positive pregnant women have been found to be infected.

Thus, the transition probabilities of vertical transmission will be developed based on the modes of vertical transmission (utero, intrapartum and postnatal). The possible transition composition of the stochastic matrix is illustrated in the form of Table 3.4 below.

Table 3.4 HIV/AIDS Transmission through vertical transmission

Recipients			
Transmitters	Utero (U)	Intrapartum (I)	Postnatal (P)
UTERO (U)	P (U, U)	P (U, I)	P (U, P)
INRAPARTUM (I)	P (I, U)	P (I, I)	P (I, P)
POST NATAL (P)	P (P, U)	P (P, I)	P (P, P)

All the infection possibilities in the nine cells of Table 3.4 are for utero (U), intrapartum (I) and postnatal (P) probabilities. The interpretation is that there exists a possibility of the child (recipient) contacting HIV/AIDS virus from an HIV/AIDS infected mother before birth (utero), during birth (intrapartum) or after birth (postnatal) as indicated in Table 3.4 above.

However to develop the transition probabilities for Table 3.4, data was collected from the HIV/AIDS cases reported to ABUTH for 20 years (1993 –2009), on the distribution of mode of vertical transmission of HIV/AIDS from mother to child. Hence, the utero, intrapartum and postnatal transitions were summed up to obtained the totals for ten years as obtained from the given data. This is tabulated and illustrated as shown in Table 3.5 below.

Table 3.5 Cumulative reported cases for HIV/AIDS through vertical transmission

Recipients			
Transmitters	Utero (U)	Intrapartum (I)	Postnatal (P)	Total
Utero (U)	24	41	4	69
Inrapartum (I)	0	43	4	47
Post Natal (P)	0	0	15	15
Total	24	84	23	131

One can now develop a three that is infected stage model from table 3.5, where state I denotes utero (U) transmitters, state II denotes intrapartum (I) transmitters, and state III denotes post natal (P) transmitters, To compute the probability of transition from state I to state I, we observed from Table 3.5, that 24 out of 69 or 35% of children (recipients) are infected before birth (utero infection), while the transition from state I to state II, is 41 out of 69 or 59% of children (recipients) are infected during birth (intrapartum infection) and the transition from state I to state III is 4 out of 69 or 6% of children (recipients) infected after birth (postnatal infection). Similarly, the transition from state II to state I is 0 out of 47 or none of the children are infected in the utero at this stage, while the transition from state II to state II is 43 out of 47 or 91% of children are infected at intrapartum and the transition from state II to state III is 4 out of 47 or 9% of children are infected at the postnatal stage. Finally, the transition from state III to state I is 0, while the transition from state III to state II is 0 and the transition from state III to state III is 15 out of 15 or the possibility of children infected at the postnatal stage is 100%.

Our probability transition matrix is then given by:

$$P = \begin{bmatrix} 0.35 & 0.59 & 0.06 \\ 0.0 & 0.91 & 0.09 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

The diagram of the transition matrix is as shown below:
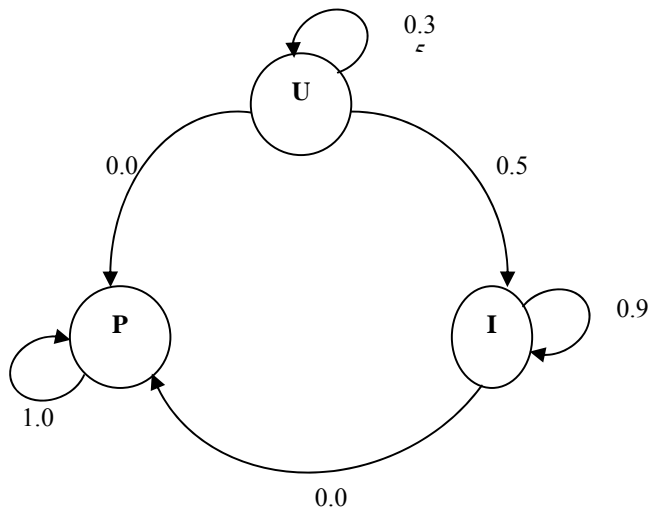
Fig 2.3 Transition Diagram for contacting HIV/AIDS through vertical transmission.

3.3 Probability State Vectors

Probability state vector has been defined as a vector of state probabilities. Suppose that, at some arbitrary time, the probability that the system is in state a_i is P_i then, these probabilities are denoted by the probability vector $P = (P_1, P_2, P_3…,P_m)$ which is called the probability distribution of the system at that time t. In particular, $P^{(0)}$ is taken to be

$$P^{(0)} = (P_1^{(0)}, P_2^{(0)}, P_3^{(0)}, …, P_m^{(0)})$$

Which denotes the initial probability distribution that is the distribution when the process begins, let

$$P^{(n)} = (P_1^{(n)}, P_2^{(n)}, P_3^{(n)}, …, P_m^{(n)})$$

denote the nth step probability distribution, i.e. the distribution after the first n^{th} steps.

If a system is known to start from a particular state, then it is assigned the value one in the vector, while others are assigned the value zero. In this project study, if it is known that the transmitter of the HIV/AIDS virus is of infected blood group A, and then the probability state vector will be (1, 0, 0, 0). If however, the transmitter is of blood group B, the vector then will be (0, 1, 0, 0).

The initial probability state vector can also be based on a survey of a sex type in an environment. It can be that a survey is taken to obtain the percentage of women infected with HIV

virus and can be obtained like (0.6, 0.4) indicating that 60% of women and 40% of men are infected respectively.

The transmitter is of blood group A and the recipient has the state vector $V = (V_A, V_B, V_{AB}, V_0)$, to find the next probability state vector, we multiply the probability state vector with the transition matrix. For example, if the initial state vector is (1,0,0) for the case of vertical transmission from infected mother to her child before birth, then the next probability vector will be.

$$(1,0,0)\begin{bmatrix} 0.35 & 0.59 & 0.06 \\ 0.0 & 0.91 & 0.09 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} = (0.35, 0.59, 0.06)$$

If, however, the state vector was found to be (0.35, 0.59, 0.06), one will have the next probability state vector calculated as

$$(0.35, 0.59, 0.06)\begin{bmatrix} 0.35 & 0.59 & 0.06 \\ 0.0 & 0.91 & 0.09 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} = (0.1225, 0.7434, 0.1341).$$

3.4 Transition Stages

Powers of the transition matrix will be studied to find out its behaviour after a number of years from the start state (transmitters). For instance, using the transition matrix

$$P = \begin{bmatrix} 0.360 & 0.64 \\ 1.0 & 0.0 \end{bmatrix}, \qquad P^2 = \begin{bmatrix} 0.7696 & 0.2304 \\ 0.3600 & 0.6400 \end{bmatrix}$$

$P^3 = P^2 \cdot P$. The matrix P is regular. If taking the powers is continued like this, it could be observed that at each power of the transition matrix, the matrix remains regular. This means that the powers of the transition matrix will continue to result in entries all assuming positive values. The sequence $P, P^2, P^3…$ of powers of P approaches the matrix T whose rows are each a fixed-point t.

4.1 RESULTS AND DISCUSSION

This section discusses the results so far obtained in this project work. This includes the movement pattern of the probability state vectors for the various modes of transmission of the HIV/AIDS virus from one year to the other. The movement pattern of the transition matrix of a particular mode of transmission from one year to another was also considered.

4.2 Movement Pattern of the Probability State Vector (PSV)

The transition matrices obtained in section 3.2, are not difficult to derive. And the explanations were elusively treated by the use of the coordinate of a lattice. If it is known to us that a male or a female are infected with the HIV/AIDS virus in a given population and that he or she had sexual relationship with another man or woman, then the transition matrix can be determined as shown in section 3.2. The initial probability vector of this mode of transmission is determined by arbitrary assignment or according to some criteria as discussed in the previous chapter.

4.2.1 The PSV Movement Pattern for the Mode of Transmission of the HIV/AIDS Virus through Sexual contact

Given that a transmitter or an infected person of the HIV/AIDS Virus through sexual contact mode of transmission is known to be a homosexual, or a heterosexual and that the initial state vector of the recipient or a susceptible person in a given population is (1, 0), after the first year, the chances that their new recipients is a male or female are respectively (0.36, 0.64). If the new recipient now a second transmitter has sexual relationship as the first transmitter, without taking any safety measures such as condom use, then the probabilities of their recipients being male or female are respectively (0.7696, 0.2304).

Continuing in this manner, we will observe that after 29 years, the probabilities of the recipients in terms of male and female will be (0.60975516312, 0.39024483689). The equilibrium is only attained at the 62nd year, where the state vector becomes (0.60975609758, 0.39024390245). At this stage (the 62nd year), it shows that irrespective of a recipient being male or female at the initial stage, the chances that a male will be infected is 0.60975609758 as long as the transmitter continue to remain infected without treating the infection or without taking safety measures during sexual intercourse with the recipient. The above also implies that the chance that a female will be infected is 0.3902439024562 after 62 years. This means that, the rate of contacting the HIV/AIDS virus is more prevalent in the male recipients than the female recipients in this population.

The curves of figure 4.2 and figure 4.1 shows the pattern of attainment of a steady state for the mode of transmission of HIV/AIDS through sexual intercourse, when a transmitter is a male and when a transmitter is a female. Figure 4.1 shows how the male transmitter probability moves steadily from 0.36 through 0.7696, of the probability at the 2nd year. The probability also decreases from 0.7696 to about 0.507456 at the 3rd year. After 19 years, the probability begins to move steadily from 0.60967504708 to converge after the 62nd year with probability 0.60975609758. That is, after a good number of years, the chances that a male transmitter will infect every recipient in that given population if there is sexual intercourse between them without taking any precaution is 0.60975609758.

The same explanation follows with the case when a transmitter is a female. Figure 4.1 shows how the probability moves steadily from 0.64, up and down through 0.39032495293 the probability at the 19th year and converges after 62 years with probability 0.39024390245.

4.2.2. The PSV Movement Pattern for the Mode of Transmission of the HIV/AIDS Virus through Vertical Transmission

An infected mother here represents a transmitter and she is fixed to be a transmitter from year to year and the offspring's or recipients the initial state vector indicating the three stages of contacting the HIV/AIDS virus through vertical transmission from an infected mother to the child is (1, 0, 0, 0). The state after the first year which indicates utero, intrapartum and postnatal stages of contacting infection at utero, intrapartum or postnatal are respectively (0.35, 0.59, 0.06). If we proceed like this, we will after 20 years attain the probabilities (0.00002758547, 0.41024863261, 0.58972378192). After 17 years, we will have the probabilities as (0.00000001775, 0.21201550164, 0.78798448061). It attains equilibrium after 85 years with probabilities (0.0000, 0.00031639071, 0.99968360929). This can be seen in the output section. This also indicates that the chances of child contacting HIV/AIDS from an infected mother are higher at the postnatal level with a possibility of about 0.99.

The curves in figure 4.1 through 4.5 also show the attainment of a steady state for the various stages of transmitting HIV/AIDS. Figure 4.3 shows the movement pattern of contracting HIV/AIDS at utero, which is before birth, the probability gradually moves from 0.35 after the first year, 0.005252218750 after 5 years, and to 0.000000000 after 20 years. After 85 years, it is observed that the probability decreases slowly to 0.0000000000, where a state of equilibrium has been attained. This shows that the rate of contracting HIV/AIDS from an infected mother decreases slowly at utero (before birth) as the year's increases.

Similarly, the curve in fig. 4.4 exhibits the movement pattern of the probability of a transmitter infecting the recipient with HIV/AIDS at intrapartum (during birth). The curve illustrates how equilibrium is attained as we progress along the years. Figure 4.4 shows the case whereby the probabilities of 0.59 from the first year decreases steadily to 0.00031631971 through 85 years equilibrium is attained. Figure 4.5 also gives the picture of how the probability of the transmitter (infected mother transmit the HIV/AIDS virus to a recipient (child) at postnatal (after birth). The probability movement pattern begins to converge at the 80th year with probability 0.99944284585. The movement pattern of the curve attains a state of equilibrium at the 85th year with probability 0.99963360929. The curves in figure 4.3 through figure 4.5 shows that transmission of HIV/AIDS from infected mother to child is more prevalent during postnatal.

4.2.3 Power of Transition Matrix for the Mode of Transmission of the HIV/AIDS Virus through Sexual Contact

The matrix was obtained of the situation when a transmitter (infected person) transmits the HIV/AIDS virus to a recipient (susceptible person) through sexual intercourse in a given population from one year to another as (see section 3.2 in chapter three).

$$\begin{bmatrix} 0.3600 & 0.6400 \\ 1.0000 & 0.0000 \end{bmatrix}$$

If we take the powers of this matrix, we will observe a moment pattern that gradually attains equilibrium or stability. After 1 year the matrix is

$$\begin{bmatrix} 0.7696 & 0.2304 \\ 0.3600 & 0.6400 \end{bmatrix},$$

After two years we will have,

$$\begin{bmatrix} 0.7456 & 0.492544 \\ 0.7696 & 0.230400 \end{bmatrix},$$

After three years we will have,

$$\begin{bmatrix} 0.67522816 & 0.32477184 \\ 0.50745600 & 0.49254400 \end{bmatrix}$$

Stability is however attained after 61 years with matrix

$$\begin{bmatrix} 0.6097560976 & 0.3902439024 \\ 0.6097560976 & 0.3902439024 \end{bmatrix}$$

Following the above trend from year 1 to year 61, we observe that during the first year, the prevalence of the disease in the homosexual group increased to 77%, while it decreases from 100% to 36% in the heterosexual group where women infect the men. Similarly, in the heterosexual group where men infect the women, the rate of spread reduces from 64% to about 23% in a given population.

In the second year the matrix also changes in its entries. Here the rate of infection possibility in the homosexual group reduces from about 77% to about 51% and that of the heterosexual group, with the men infecting the women increases from 23% to about 49%. Similarly, the heterosexual group with women infecting the men increases from 36% to about 77%. The trend from year 1 through year 3 seems to fluctuate since from the first year running through the seven year we observe that there is a decrease/increase in the dynamics of the spread of the HIV/AIDS virus, because, there is a sharp decrease and increase for both the heterosexual groups with men infecting the women or with women infecting the men, as well as the homosexual groups. As the trend continues from year to year we observed that equilibrium was attained after 61 years. This means that the dynamics of the spread of infection through the sexual intercourse mode of transmission remains stable after 61 years. Hence, the infection possibilities for each of the cells can be used to predict the prevalence of the disease in a given population.

The stable possibility of contacting infection in the homosexual cell is 0.6097560976; while for the heterosexual cell with men infecting the women is 0.3902439024. This means that the stable possibility of infection can be use to forecast the number of persons that can be infected with the HIV/AIDS virus for both the homosexual and heterosexual cells with men infecting the women in a given population. For example, in a given population of 1,000 susceptible persons, we will observe that if nothing is done to combat the spread of the HIV/AIDS virus through sexual contact, about 610 persons will be infected after 61 years through homosexual relationships and about 390 persons will be infected through heterosexual relationships with the men infecting the women. The trend will follow suit for the heterosexual cell with the women infecting the men. On the whole the matrix at equilibrium shows that the possibility of infection for the homosexual cell and heterosexual cell (women infecting the men) have equal probabilities, while for the heterosexual cell (men infecting the women) have probability of 0.6097560976

and 0.3902439024 respectively. The matrix at equilibrium can be represented as seen in the Table 4.1

Table 4.1 *The Equilibrium stage when a transmitter transmits the HIV/AIDS virus to a recipient, through sexual intercourse.*

Recipient		
Transmitter	**Male (M)**	**Female (F)**
Male (M)	0.6097560976	0.3902439024
Female (F)	0.6097560976	0.3902439024

The matrix at equilibrium is regular, since all entries are positive. The matrix has attained a steady equilibrium, since multiplying this matrix with transition matrix still result in the same matrix. The stage in which the matrix passes through before reaching equilibrium from year to year can be seen clearly in output section.

4.2.4 Power of Transition Matrix for the Mode of Transmission of the HIV/AIDS Virus through Blood Transfusion

The matrix obtained when a transmitter transmits the HIV/AIDS virus to a recipient through blood transfusion, is represented as shown below:

$$P = \begin{pmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1,0000 & 0.0000 \\ 0.1900 & 0.2500 & 0.2300 & 0.2300 \end{pmatrix}$$

After the first year the matrix becomes,

$$\begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 1.000 & 0.000 \\ 0.190 & 0.250 & 0.230 & 0.230 \end{pmatrix}$$

Table 4.2 *The Equilibrium stage when a transmitter transmits the HIV/AIDS virus, through blood transfusion.*

Recipient				
Transmitter	**Group A**	**Group B**	**Group AB**	**Group O**
Group A	1	0	0	0
Group B	0	1	0	0
Group AB	0	0	1	0
Group O	0.2527	0.3325	0.3059	0.1089

4.1.5. Power of Transition Matrix for the Mode of Transmission of HIV/AIDS through Vertical Transmission

The matrix for the vertical transmission of the HIV/AIDS virus from a HIV infected mother to her child was obtained to be:

$$P = \begin{pmatrix} 0.3500 & 0.5900 & 0.0600 \\ 0.0000 & 0.9100 & 0.0900 \\ 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

If we take the powers of this matrix one will observe that the matrix attains stability after one year. The matrix after 1 year is

$$\begin{pmatrix} 0.1225 & 0.7434 & 0.1341 \\ 0.0000 & 0.8281 & 0.1719 \\ 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

The matrix has attained a steady state after 1 year since, multiplying this matrix result in the same matrix. What this means is that irrespective of the start state of the transmitter, the probabilities of the off springs being infected at various stages of vertical transmission remains after 1 year. The movement pattern of the transition matrix shows that the dynamics of the spread of the disease is more prevalent at the postnatal stage. The prevalence of infection at the utero stage and at the intrapatum stage, have respective probabilities at utero and intrapatum levels as 0.1225 and 0.7434. Hence, contacting infection is less prevalent at utero, compared to the intrapatum level. This is illustrated more clearly in Appendix V. However, the matrix at equilibrium can be represented as seen in table 4.3.

Table 4.3 *Equilibrium stage when a transmitter transmits the HIV/AIDS virus through vertical transmission.*

	Recipient		
Transmitter	**Utero**	**Intrapartum**	**Postnatal**
Utero	0.1225	0.7434	0.1341
Intrapartum	0.000	0.8281	0.1719
Postnatal	0.00	0.00	1.00

4.3 Curve Movement Pattern of Probability State Vector (PSV)

The computer was also used to obtained diagrams of the curve movement patterns of the state vector from year to year.

For each of the modes of transmission of the HIV/AIDS virus from the infected persons (transmitter) to a susceptible person (recipient), one curve is used to illustrate each case for only one probability state vector, The curve movement patterns of the infected person (transmitter) to a recipient (susceptible person) are given in figures 4.1 to figure 4.5 below.
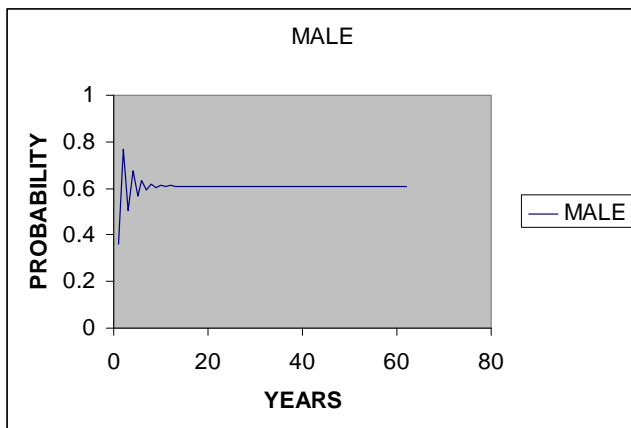
Fig. 4.1 Curve Movement Pattern of the mode of transmission of HIV/AIDS through sexual contact, when the transmitter is a male; IPSV (1,0)
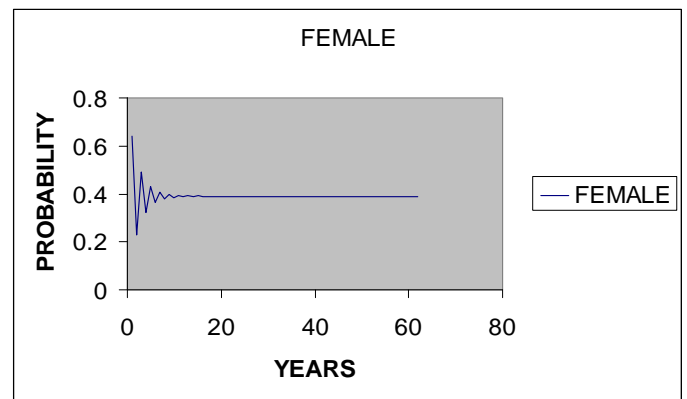
Fig. 4.2 Curve Movement Pattern of the mode of transmission of HIV/AIDS through sexual contact, when the transmitter is a female; IPSV (1,0)
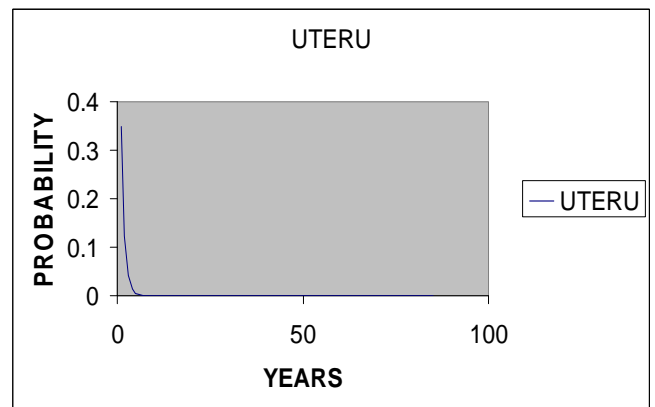
Fig. 4.3 Curve Movement Pattern of the mode of transmission of HIV/AIDS through vertical transmission at Uterus; IPSV (1,0,0)
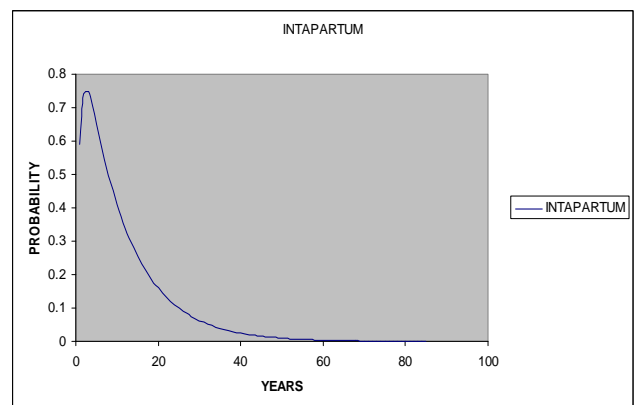
Fig. 4.4 Curve Movement Pattern of the mode of transmission of HIV/AIDS through vertical transmission at Intrapartum; IPSV (1,0,0)
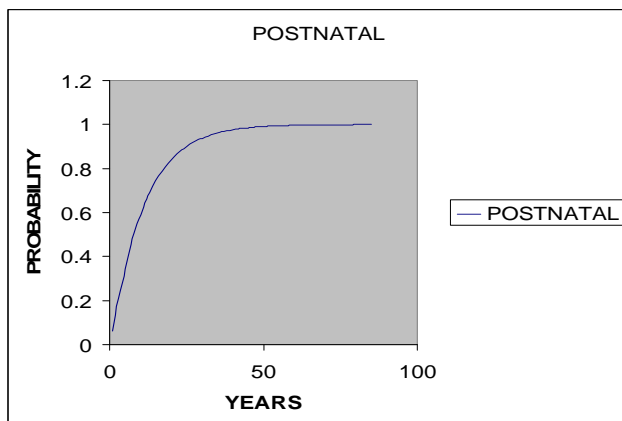
Fig. 4.5 Curve Movement Pattern of the mode of transmission of HIV/AIDS through vertical transmission at Postnatal; IPSV (1,0,0)

5.1 Discussion

Stochastic processes for the epidemic is required to produce projections of the epidemic such as the spread and control of the virus, upon which public policy and plans for providing and financing health care services can be based.

We explained how the transition matrices of the various modes of transmission of HIV/AIDS virus were developed. When the start state of the recipient (susceptible person) is known, then the initial probability state vector is known. Possible probabilities of infected recipients from transmitters (infected persons who transmit the HIV/AIDS virus) were determined by taking the product of the probability state vector with the transition matrix. It is observed from the result obtain that in chapter four, of the product of the PSVs with the transition matrices, as one moves from one year to another, made the probability state vectors converge to a steady vector.

5.2 Conclusion

The patterns of the transition matrices was observed and the following trends were obtained; for the mode of transmission of the HIV/AIDS virus through sexual intercourse, its transition matrix, of 61 years are required for equilibrium to be attained, while for the mode of transmission through blood transfusion, its transition matrix, a year only is required to attain a state of equilibrium and for the mode of transmission of HIV/AIDS virus through vertical transmission from HIV infected mother to her child, its transition matrix, a year is required for equilibrium to be attained.

From this study, people should insist on screening of blood to ensure that it is free from the virus, before embarking on transfusion. They should also do it to the person they intend to marry. People should also insist on condom use before sex and use of sterilized objects that are capable of cutting or injecting the body.

For every pregnant woman, scanning should be undergone before or during antenatal and proper action should be taken where necessary. Prevention is better than cure.

However, it seems likely that the new treatments and other preventive measures may have profound implications for future predictions in a given population, but considerable uncertainty remains about their longer-term efficacy. The results suggest that we need to extend our methodology to incorporate the apparent effects of new treatments, preventive measures and to allow for other causes of death in these risk groups.

REFERENCE

Cooley P. C., Hamill P. C. and Myers L. E. (1993) A linked risk group model for investigating the spread of HIV, *Journal of Mathematical computation modeling vol.18, No.12, PP 85 – 102.*

Maki D.P (1989). **Finite Mathematics**. M C Graw-Hill Book Company, New York.

Nigeria Bulletin of Epidemiology. *A publication of the Epidemiological Division Disease Control and International Health Department. FMHHS.* Vol. 2, No. 2(1992

Tan W.Y and Byers R.H (1993) A Stochastic Models of the HIV Epidemic and the HIV Infection Distribution in a Homosexual Population, math. *Biosciences 113,115-143*

Otubu R.N (1998) Need to emphasize vertical transmission in HIV/AIDS strategy " *Journal of Nigeria medical association vol. 8,pp.88-90"*

Norman T.J (1961) The Mathematical Theory of InfectioUs Diseases and its Application Oxford, University Press**.**

Webpage Classification based on URL Features and Features of Sibling Pages

Sara Meshkizadeh,
Department of Computer engineering,
Science and Research branch,Islamic Azad University(IAU) ,
Khouzestan,Iran
Sara_meshkizadeh@yahoo.com

Dr.Amir masoud rahmani,
Department of Computer engineering,
Science and Research branch,Islamic Azad University(IAU) ,
Tehran,Iran
rahmani@srbiau.ac.ir

Dr.Mashallah Abbasi Dezfuli
Department of Computer engineering,
Science and Research branch,Islamic Azad University(IAU) ,
Khouzestan,Iran,
Abbasi_masha@yahoo.com

***Abstract*:Webpage classification plays an important role in information organization and retrieval. It involves assignment of one webpage to one or more than one predetermined categories. The uncontrolled features of web content implies that more work is required for webpage classification compared with traditional text classification. The interconnected nature of hyper text, however, carries some features which contribute to the process, for example URL features of a webpage. This study illustrates that using such features along with features of sibling pages, i.e. pages from the same sibling, as well as Bayesian algorithm for combining the results of these features, it would be possible to improve the accuracy of webpage classification based on this algorithm.**
Keywords: classification, hyper text, URL, sibling pages, Bayesian algorithm

I. INTRODUCTION

Traditional classification is supposed to be a instructor-based teaching where a set of labeled data can be used for teaching a classifier to be employed for future classification. Webpage classification is distinct from traditional on in a number of aspects. First, traditional textual classification is usually performed on the structured documents written based on a fixed style (e.g. news articles), whereas webpage content is far from such characteristics. Second, web pages are documents with HTML structure which might be translated for the user visually. Classification plays a significant role in information management and retrieval task. On the web, webpage classification is essential for focused crawling, web directories, topic specific web link, contextual advertising, and topical structure. It can be of great help in increasing search quality on the web. Using URL related information, [1] has achieved 48% accuracy in classification. [2] uses URL information as well as

webpage pictures to provide a classification with tree structure. [3] involves combination of parent page information with HTML features to arrive at classification. [4] reaches 80% accuracy through a combination of web mining information. Combining some features of HTML and URL and in sum 9 features, classification is made with 80% accuracy. This study introduces a novel method for webpage classification as it enhances existing algorithms. This method combines three different features of webpage URL, and it also employs information from neighboring pages with the same sibling, i.e. sibling pages to increase classification accuracy.

In this paper, in Section 2 related concepts are described. Section 3 introduces the proposed algorithm in detail. Section 4 evaluates the suggested algorithm, and Section 5 provides conclusions and insights towards future work.

II. RELATED CONCEPTS

A. Selecting Database

As webpage classification is usually taken as a supervised learning, there is a need for classified samples for learning. In addition, some samples are also required to test classification for classifier evaluation. Manual labeling involves something more than human, therefore, some available web directories are used for more research. One of them which is employed more is OPD[12]. In this database, 4519050 various websites are classified by 84430 classification editors. This research has been performed on universities, shopping, forums, FAQ (frequently asked questions) of ODP.

B. Feature Extraction from Webpage URL

The contextual content is the most important feature available on the webpage. However, taking diverse range of parasites on webpage into account, using bag of words directly may does not lead to higher

efficiency. The researchers have proposed a number of methods for better application of contextual characteristics. Feature selection is a popular one. In addition to features of HTML tags, a webpage can be classified based on its own URL. URLs are highly effective in classification. First, a URL is easily retrievable, and each URL is limited to one webpage, and each webpage has one special URL. Second, if this method is solely employed, classification of one webpage based on its URL causes download removal of the whole page. This tends to be an appropriate method for classifying pages which are not existing or their download is impossible, or time/space is critical for example in realtime classifications.

C. Bayesian Algorithm

Bayesian inference is provides a probability method for inference. This method is built on the hypothesis that the considered values follow a probable distribution, and that optimal decisions can be made with an eye to inference on the probabilities as well as observed data. As this method is a quantitative one for weighing evidences which support different hypotheses, it is of great importance in machine learning. Bayesian inference provides a direct method for dealing with probabilities for learning algorithms, and it also creates a framework for analyzing performance of algorithms which are not directly related to probabilities. In many cases, the problem is finding the best hypothesis in hypothesis space H with available D learning data. One method to express the best hypothesis is that we claim we are looking for the most probable hypothesis with D data in addition to initial data on prior probabilities H. Bayesian theorem is also a direct method for calculating these probabilities.

To define Bayesian theorem, P(h) is used to express initial probability which maintains h hypothesis is true, earlier than observing learning data. P(h) is usually called prior probability, and it expresses any prior knowledge which states on the chance of correctness of hypothesis h. If there is no initial knowledge on hypotheses, we can assign a similar probability to the whole hypotheses space H. Likewise, P(h) is similarly used for expressing prior probability where D data are observed. In other words, probability of observing D in the case of there is no knowledge on correctness of hypotheses. P(D|h) is employed to express probability D in a space where hypothesis h is true. In machine learning we look for P(D|h), i.e. probability of correctness of hypothesis h in the case of observing D learning data. P(D|h) is called post probability, as it expresses our confidence of hypothesis h after observing D data.

Bayesian theorem is the main building block of Bayesian learning, as it provides a method for calculating post probability P(D|h) based on P(h) along with P(D) and P(D|h).

$$P(h \mid D)= \frac{P(D \mid h)\, P(h)}{P(D)} \quad (1)$$

As expected, it can be seen that P(h|D) increases with the increase of P(h), and P(D|h). Therefore, it is reasonable that P(D|h) decreases with the increase of P(D), because with higher probability of occurrence P(D) which is independent from h, fewer evidence of D are available to support h. In many learning scenarios, the learner considers a set of hypotheses H, and it is interested in finding the hypothesis h∈H which is the most probable one or at least one of the most probable ones. Any hypothesis which carries such feature is called Maximum a posteriori, MAP. Using Bayesian theorem, it is possible to find MAP hypothesis to calculate posteriori probability of each candidate. In other words, HMAP is a hypothesis that

$$
\begin{aligned}
\text{HMAP} &= \text{argmax } hj \in H \ \ P(hj \mid Di) \\
&= \text{argmax } hj \in H(\ P(Di \mid hj)\ P(hj)\ / P(D)) \quad (2) \\
&= \text{argmax } hj \in H\ P(Di \mid hj)\ P(hj)
\end{aligned}
$$

Be noted that P(D) is deleted at the final step, as its calculation is independent from h, and it is always a constant. However, for all probable states on different features of a problem, this theorem can be generalized for all existing probabilities, as for all values of feature D1,D2…Dn:

$$H = \text{argmax } hj \in H\ \ P(hj)\ \Pi\ P(Di \mid hj) \quad (3)$$

And P(Di|hj) is calculated as follow:

$$P(Di \mid hj)= \frac{n\_c + mp}{n+m} \quad (4)$$

Where:
n= number of examples where h= hj
nc = number of examples where samples D= Di and h= hj
p= initial estimation for P(Di|hj)
m= size of sample space

D. Using Features of Neighbor Webpage

Although URL addresses of webpage contain useful features, they may miss, misunderstood, or not recognized for some reasons in a special URL. For example, some pages do have brief addresses or concise contextual content. In such cases, it would be difficult for classifiers to decide on the accuracy according to URL features of the webpage. To solve the problem, these features might be extracted from neighbor pages which are related to the webpage under classification. In this study, pages with the same sibling, i.e. sibling pages are employed to arrive at more accuracy in classification.

III. Proposed Algorithm
A. *Pre-processing of Web Content*

In most studies, pre-processing is performed prior to feeding the web content to the classifier. In the proposed method here, only URL address and page Title are required for implementing classification algorithm. First, the URL address and page Title are pre-processed, i.e. words shorter than 3 characters, numbers, conjunctions and prepositions stored in a table called Stopwords are removed from this content. Using the function Porter Stemmer [13], the useful words are changed to their stems (to avoid data redundancy), and then they are stored in the relevant table along with a value as the frequency of word in the data bank of the program.

B. *Using Extracted Features from Webpage URL*

1) *First Feature*

URL do have appropriate features for classification. Two sets of features which are easily extracted from URL pages are Postfix and Directory. The general format of a postfix is usually as Abbreviation or Abbreviation. Abbreviation. For example, .edu or ac.ir or .ac.uk indicate pages related to universities or academic websites. The general format of a Directory feature is mostly like Word(Abbreviation)slash. For example, a directory named FAQ or Forum represented as /Faq/ or /Forum/ can represent the relation between the current page to the relevant class. These features are put in Table 1.

Table 1: Features of URL Addresses

URL feature	Specification	number
Postfix : .edu,.ac.ir	**To find class of pages based on Stem URL Address**	1
Directory : Forums/,Faq/	**To find class of pages based on rest of Stem URL Address**	2

2) *Second Feature*

The second feature important in webpage URL is attention to the domains which have been observed by the system and further their correct class is identified. To do so, once the class of a webpage is determined at the last step and its accuracy is confirmed by the user, the page's address is registered in the URL table along with the correct class. Afterwards, if a webpage with the same domain is given to the system, the system can recognize more simply and quickly based on similarity in addresses. Take for example the address www.aut.ac.ir/sites/e-shopping/raja.ir whose domain is recognized as shopping, thereby it is registered under shopping class at the URL table.

If a webpage such as www.aut.ac.ir/sites/e-shopping/iranair.ir is given to the system as a test, due to the domain www.aut.ac.ir/sites/e-shopping/ in the table, the system quickly and simply assigns shopping class to the second address. It should be noted that using domain similarity of webpage URL is a new idea which has not been taken into account in webpage classification.

3) *Third Feature*

To enhance the efficiency of the proposed algorithm another method based on URL address can also be employed. This method involves expanding URL of a webpage to use existing elements better. Forexample,considering http://www.washington.edu/news/nytimes , and also page title, i.e. NewYorkTimes it is easily understood that the address refers to NewYorkTimes news database. The machine, however, misses the point. To solve the problem and to employ human guesses in the procedure, with an eye to the function used in [1] but with some changes in definition and usage, a likelihood function is presented. Table 2 illustrates the method. The likelihood is used to compare URL's token letter by letter, and similar word in the page Title.

Table2. Likelihood Function

Rank	Condition	number
2	First letter of URL token is the same as first letter of similar word at page Title	1
1	First letter of URL token is the same another letter of similar word at page Title	2
1	A letter of URL token is the same as a letter of similar word at page Title	3
2	Last letter of URL token is the same as last letter of similar word at page Title	4
0	Ignoring a character of URL token	5

For example consider the news website Newyorktimes with the address http://www.nytimes.com. Now take the title "The New York Times-Breaking News, World News & Multimedia". The above-mentioned function calculates likelihood for a condition where URL token, Nytimes and similar word at page Title is Newyorktimes as bellow:

(Condition1)$\rightarrow$N(Condition5)$\rightarrow$E(Condition5)$\rightarrow$ W(Condition1)$\rightarrow$Y(Condition5)$\rightarrow$O(Condition5)$\rightarrow$ R(Condition5)$\rightarrow$K(Condition1)$\rightarrow$T(Condition3)$\rightarrow$ I(Condition 3)$\rightarrow$M(Condition 3)$\rightarrow$E(Condition 4)$\rightarrow$S

In this case, the word Newyorktimes receives score 11 from words at the page Title which is the highest score out of other tokens, and according to the likelihood

function, thereby in future if Nytimes is seen in test samples, it is replaced with Newyorktimes. To work with the third feature, the system runs the likelihood function on tokens of URL and the page Title. In the case of expanding existing token in URL, instead of the former token, an expanded token is used for classification.

C. Webpage Classification

1) Training Step

In this step, about 2000 pages of considered classes in ODP database are used for system learning. As such, the webpage URL address, the page's Title along with the category is fed to the system. In this step, useful words from considered features of the page URL are extracted, and are stored in the table related to the considered class of the program information bank along with the frequency of each word in each feature as the value of that word in that feature. In this step, information of pages in each category is stored in the table related to the same category. So far we have been dealing with learning the system and feeding the useful information.

2) Test Step

In this step, to test the system a number of different pages in Training step is used. To do so, URL address and the page Title are given to the system. The system, based on the first feature analyses the possibility of recognizing the category according to the address and checking features of Postfix and Directory. If it fails to find a previously observed Postfix or Directory, it refers to the second feature, and compares the page address with all addresses observed so far. After this step, it is obvious that accuracy of this algorithm increases as the number of learned pages considerably increases. If the address of this page is not similar to existing addresses of the bank, the algorithm studies the third feature. It must be stated that the first priority of classification system is the first feature, followed by the second feature and finally the third one. To analyze the third feature, the system performs the likelihood function on existing tokens in URL and page Title. If the existing tokens in URL are expanded, instead of former token an expanded expression is employed for classification. Afterwards, the frequency of each word is calculated as frequency feature. Then using Bayesian rule, posteriori probability $P(h|D)$ from priori probability $P(h)$ along with $P(D)$ and $P(D/h)$ according to frequency of each word on the page is calculated in order to predict the possibility of belonging the page to the category where the highest number of features has been occurred. After calculating the probability for all extracted words of one page for all four categories, the category which has the highest probability for all words is recognized as the main category and is announced to the user.

3) Neighbor Effect Step

Due to the concise nature of most addresses and also Title of many webpages, the classification based on just features extracted from a webpage has been performed with 72.8% accuracy as shown at Table 3. To improve classification accuracy, it was decided to use extracted information from neighbor pages. As such, URL address and Title of 500 neighbor pages related to Test step pages have been employed to increase classification accuracy. As illustrated in Table 3, the accuracy average raised to 85.5% that implies the emphasis on neighbor pages is helpful. As described earlier, the system first analyses by the first feature. If it fails it refers to the second feature, and ultimately the third feature is used for classification. The accuracy of averages of algorithm is shown in Table 3. The result of combining former features with sibling pages features and their influence can be observed.

Table 3. Results of Algorithm Accuracy based on different Features

feature	Shopping	University	Forums	FAQ	F-measure
Based on feature 1	%48.4	%65.8	%61.4	%63.2	%59.7
Based on feature 2	%76.2	%74.5	%70.7	%69.8	%72.8
Based on feature 3	%66.3	%61.5	%68.4	%64.6	%65.2
Based on combining former features and sibling pages features	%84.5	%85.6	%84.7	%88.4	%85.8

Figure 1 illustrates the results of algorithm evaluation for different categories and based on different features.
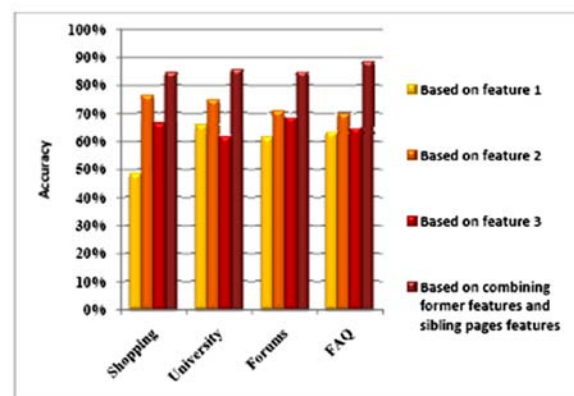
Figure 1. Results of Algorithm Evaluation for different Categories and Features

IV.   Evaluation of Proposed Algorithm

In this study, to evaluate the proposed algorithm, 500 different pages with training and test steps, and categories downloaded from ODP such as universities, shopping, forums, and FAQ have been employed. To implement the algorithm, Vb.net 2005 programming environment as well as Sqlserver 2000 database on a computer with specifications cpu core 2,2.13 GHz, and 2 GB Ram, and 300 GB Hard disk which ran the operating system Windows XP service pack 3 is used. After pre-processing downloaded pages and extraction of useful features from URL address and Title of pages, and calculating the probability of belonging of the webpage to each one of categories using Bayesian algorithm, and then calculating the above probability based on neighbor pages information, the category with the highest probability is recognized as main page category and it is announced to the user. If the category is recognized correctly, all its extracted features are added to the table of program's information bank, and is added as a correct case to the statistics of correct recognition, parameter a from calculation criterion of algorithm's general accuracy is added too. Otherwise, the correct category is recognized by the user, and the extracted information is added from the page to the correct table. Furthermore, one case to wrongly recognized cases from the first category, i.e. parameter b from calculation criterion of algorithm's general accuracy, and also one case to cases which is wrongly ignored in the correct category, i.e. parameter c from calculation criterion of algorithm's general accuracy are added. To recognize the algorithm accuracy, algorithm accuracy evaluation criterion has been used as follows:

$$Precision = a / (a+b) \quad (5)$$
$$Recall = a / (a+c) \quad (6)$$
$$F\text{-measure} = 2 * Recall * Precision / (Recall + Precision) \quad (7)$$

where a= number of pages of sample examples which are correctly classified, b= number of pages of sample examples which are wrongly classifies, and c= number of sample examples which are wrongly ignored.

In this study, the proposed algorithm has been evaluated in four phases. In the first phase, the evaluation of algorithm implementation on the first feature of the webpage URL with no reference to other URL features illustrated 59.7% F-measure accuracy. In the second phase, using the second selected feature raised algorithm accuracy to 72.8%. The second feature, i.e. domain likelihood in the webpage address is a new feature not employed in similar studies and higher algorithm accuracy is an evidence to the fact. The third phase is evaluation of using likelihood function in classification which compared with [1] which shows 43% accuracy with the same feature, illustrated 65% accuracy thanks to

its combination with Bayesian algorithm, and implying that this approach waits for further work. The fourth phase is the evaluation of combination of all three proposed features related to URL and using the features of neighbor pages in classification which arrived at 85.8% accuracy in classification. Comparing this accuracy with [5] that reached 80%, or [4] that toped at 63% through Bayesian implementation, it is approved that the proposed method is more efficient and promising. The algorithm's average accuracy, F-measure is shown in different categories in Figure 2.
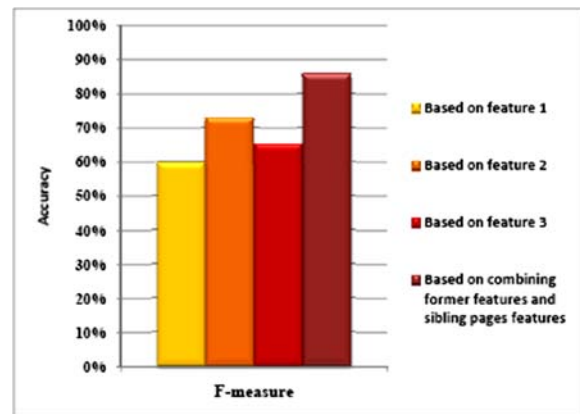
Figure 2. Algorithm Average Accuracy based on different Features

V.   Conclusion and Future Works

In this study, based on URL features, a new method is presented for webpage classification. This method involves combining three different features from pages URL as well as information of neighbor page information for higher accuracy in classification. The proposed algorithm ultimately achieved 85.8% accuracy, and it can be enhanced through combining this method with the methods based on HTML features of webpage.

VI.   Refrences

[1]- M.kan,2004,"Web page categorization without the web page", Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, ACM,OnPages : 262 - 263 .

 [2]-L.k.shih,D.r.karger, 2004,"Using URLs and Table Layout for Web Classification Task", Proceedings of the 13th international conference on World Wide Web ,ACM, OnPages : 193 - 202 .

[3]-X.qi,B.D.Davison,   2008,"classifiers   without borders : incorporating fielded text from neighboring webpages", In Proceedings of the 31 st Annual international ACM SIGIR Conference on Research & Development on Information Retrieval, Singapore, July 2008, On page(s):643-650.

[4]-S.Morales,H.Fandino,J.rodriguez,   2009,"
Hypertext Classification to filtrate information on the web", Proceedings of the 2009 Euro American

Conference on Telematics and Information Systems: New Opportunities to increase Digital Citizenship 2009, Prague, Czech Republic   June 03 - 05, 2009, Article No. 1 .

[5]-Ch.Lindemann,L.littig,   2006,"Coarse-grained classification of websites by their structural properties", Proceedings of the 8th annual ACM international workshop on Web information and data management table of contents,Arlington, Virginia, USA .OnPage(s): 35 - 42   .

[6]- Sebastiani, F.,2002," Machine learning in automated text categorization". ACM Computing Surveys (CSUR) archive,Volume 34 , Issue 1 (March 2002),On Page(s): 1 - 47 .

[7]- Chakrabarti, S, Morgan Kaufmann. 2003," Mining the Web: Discovering Knowledge from Hypertext Data",San Francisco, CA. Morgan-Kaufmann Publishers.

[8]- Mladenic, D, 1999," Text-learning and related intelligent agents: A survey", Intelligent Systems and their Applications, IEEE, Jul/Aug1999, Volume: 14, issue   4,Onpage(s):44-54.

[9]- Getoor, L, Diehl, C. 2005" Link mining: A survey", ACM SIGKDD Explorations Newsletter archive (Special Issue on LinkMining), December 2005,Volume 7 , Issue 2 ,On Page(s): 3 – 12.

[10]- Furnkrunz, J. 2005," Web mining. In The Data Mining and Knowledge Discovery Handbook", O. Maimon and L. Rokach, Eds. Springer, Berlin, Germany, On Page(s): 899–920

[11]- Choi, B. , Yao. Z, 2005," Web page classification. In Foundations and Advances in Data Mining" W. Chu and T. Y. Lin, Eds. Studies in Fuzziness and Soft Computing, vol. 180. Springer-Verlag, Berlin, Germany, On Page(s):221–274.

[12]-www.dmoz.org

[13]-http://tartarus.org/~martin/PorterStemmer/

Clustering Unstructured Data (Flat Files)

An Implementation in Text Mining Tool

Yasir Safeer[1], Atika Mustafa[2] and Anis Noor Ali[3]

Department of Computer Science

FAST – National University of Computer and Emerging Sciences

Karachi, Pakistan

[1]yasirsafeer@gmail.com, [2]atika.mustafa@nu.edu.pk, [3]anisnoorali@hotmail.com

Abstract—**With the advancement of technology and reduced storage costs, individuals and organizations are tending towards the usage of electronic media for storing textual information and documents. It is time consuming for readers to retrieve relevant information from unstructured document collection. It is easier and less time consuming to find documents from a large collection when the collection is ordered or classified by group or category. The problem of finding best such grouping is still there. This paper discusses the implementation of k-Means clustering algorithm for clustering unstructured text documents that we implemented, beginning with the representation of unstructured text and reaching the resulting set of clusters. Based on the analysis of resulting clusters for a sample set of documents, we have also proposed a technique to represent documents that can further improve the clustering result.**

Keywords—**Information Extraction (IE); Clustering, k-Means Algorithm; Document Classification; Bag-of-words; Document Matching; Document Ranking; Text Mining**

I. INTRODUCTION

Text Mining uses unstructured textual information and examines it in attempt to discover structure and implicit meanings "hidden" within the text [6]. Text mining concerns looking for patterns in unstructured text [7].

A cluster is a group of related documents, and clustering, also called unsupervised learning is the operation of grouping documents on the basis of some similarity measure, automatically without having to pre-specify categories [8]. We do not have any training data to create a classifier that has learned to group documents. Without any prior knowledge of number of groups, group size, and the type of documents, the problem of clustering appears challenging [1].

Given N documents, the clustering algorithm finds k, number of clusters and associates each text document to the cluster. The problem of clustering involves identifying number of clusters and assigning each document to one of the clusters such that the intra-documents similarity is maximum compared to inter-cluster similarity.
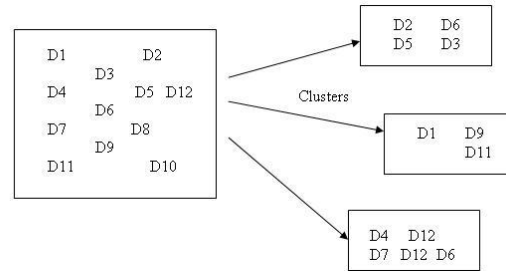
Figure 1. Document Clustering

One of the main purposes of clustering documents is to quickly locate relevant documents [1]. In the best case, the clusters relate to a goal that is similar to one that would be attempted with the extra effort of manual label assignment. In that case, the label is an answer to a useful question. For example, if a company is operating at a call center where users of their products submit problems, hoping to get a resolution of their difficulties, the queries are problem statements submitted as text. Surely, the company would like to know about the types of problems that are being submitted. Clustering can help us understand the types of problems submitted [1]. There is a lot of interest in the research of genes and proteins using public databases. Some tools capture the interaction between cells, molecules and proteins, and others extract biological facts from articles. Thousands of these facts can be analyzed for similarities and relationships [1]. Domain of the input documents used in the analysis of our implementation, discussed in the following sections, is restricted to Computer Science (CS).

II. REPRESENTATION OF UNSTRUCTURED TEXT

Before clustering algorithm is used, it is necessary to give structure to the unstructured textual document. The document is represented in the form of vector such that the words (also called features) represent dimensions of the vector and frequency of the word in document is the magnitude of the vector. i.e.

■ A Vector is of the form
$$<(t1,f1),(t2,f2),(t3,f3),....(tn,fn)>$$
where t1,t2,..,tn are the terms/words(dimension of the vector) and f1,f2,…,fn are the corresponding frequencies or magnitude of the vector components.

A few tokens with their frequencies found in the vector of the document [9] are given below:

LIST OF FEW TOKENS WITH THEIR FREQUENCY IN A DOCUMENT

Tokens	Freq.	Tokens	Freq.
oracle	77	cryptographer	6
attacker	62	terminate	5
cryptosystem	62	turing	5
problem	59	return	5
function	52	study	5
key	46	bit	4
secure	38	communication	4
encryption	27	service	3
query	18	k-bit	3
cryptology	16	plaintext	3
asymmetric	16	discrete	2
cryptography	16	connected	2
block	15	asymptotic	2
cryptographic	14	fact	2
decryption	12	heuristically	2
symmetric	12	attacked	2
compute	11	electronic	1
advance	10	identifier	1
user	8	signed	1
reduction	8	implementing	1
standard	7	solvable	1
polynomial-time	7	prime	1
code	6	computable	1
digital	6		

The algorithm of creating a document vector is given below [2]:

TABLE II.   GENERATING FEATURES FROM TOKENS

Input
Token Stream (TS), all the tokens in the document collection

Output
HS, a Hash Table of tokens with respective frequencies

Initialize:
Hash Table (HS):= empty Hash Table

for each Token in Token Stream (TS) **do**
        If Hash Table (HS) contains Token **then**
         Frequency:= value of Token in hs
         increment Frequency by 1
        else
         Frequency:=1
        enidif
        store Frequency as value of Token in Hash Table (HS)
endfor

output HS

Creating a dimension for every unique word will not be productive and will result in a vector with large number of dimensions of which not every dimension is significant in clustering. This will result in a synonym being treated as a different dimension which will reduce the accuracy while computing similarity. In order to avoid this problem, a Domain Dictionary is used which contains most of the words of Computer Science domain that are of importance. These words are organized in the form of hierarchy in which every word belongs to some category. The category in turn may belong to some other category with the exception of root level category.

Parent-category → Subcategory→Subcategory → Term(word). e.g. Databases→RDBMS→ERD→Entity

Before preparing vector for a document, the following techniques are applied on the input text.
- The noise words or stop words are excluded during the process of Tokenization.
- Stemming is performed in order to treat different forms of a word as a single feature. This is done by implementing a rule based algorithm for Inflectional Stemming [2]. This reduces the size of the vector as more than one forms of a word are mapped to a single dimension.

The following table [2] lists dictionary reduction techniques from which Local Dictionary, Stop Words and Inflectional Stemming are used.

TABLE III.   DICTIONARY REDUCTION TECHNIQUES

Local Dictionary
Stop Words
Frequent Words
Feature Selection
Token Reduction: Stemming, Synonyms

A. tf-idf Formulation And Normalization of Vector

To achieve better predictive accuracy, additional transformations have been implemented to the vector representation by using *tf-idf* formulation. The *tf-idf* formulation is used to compute weights or scores of a word. In (1), the weight $w(j)$ assigned to word j in a document is the *tf-idf* formulation, where j is the j-th word, $tf(j)$ is the frequency of word j in the document, N is the number of documents in the collection, and $df(j)$ is the number of documents in which word j appears.

Eq. (1) is called inverse document frequency (*idf*). If a word appears in many documents, its *idf* will be less compared to the word which appears in a few documents and is unique. The actual weight of a word, therefore, increases or decreases depending on *idf* and is not dependent on the term frequency alone. Because documents are of variable length, frequency information could be misleading. The *tf-idf* measure can be normalized to a unit length of a document D as described by *norm(D)* in (3) [2]. Equation (5) gives the cosine distance.

$$w(j) = tf(j) * log_2(N/df(j)) \qquad (1)$$

$$log_2(N/df(j)) = idf(j) \qquad (2)$$

$$norm(D) = \sqrt{\sum w(j)^2} \qquad (3)$$

$$\hat{D} = \vec{D}/norm(D) \qquad (4)$$

$$cosine(d1, d2) = \sum \left(w_{d1}(j) * w_{d2}(j)\right)/(norm(d1) * norm(d2)) \qquad (5)$$

e.g.
For three vectors (after removing stop-words and performing stemming),

- *Doc1 < (computer, 60), (JAVA, 30)...>*
- *Doc2 < (computer, 55), (PASCAL, 20)...>*
- *Doc3 < (graphic, 24), (Database, 99)...>*

Total Documents, *N*=3

The vectors shown above indicate that the term '*computer*' is less important compared to other terms (such as *'JAVA'* which appears in only one document out of three) for identifying groups or clusters because this term appears in more number of documents (two out of three in this case) making it less distinguishable feature for clustering. Whatever the actual frequency of the term may be, some weight must be assigned to each term depending on the importance in the given set of documents. The method used in our implementation is the *tf-idf* formulation.

In *tf-idf* formulation the frequency of term *i,* tf*(i)* is multiplied by a factor calculated using inverse-document-frequency *idf(i)* given in (2). In the example above, total number of documents is *N*=3, the term frequency of '*computer*' is $tf_{computer}$ and the number of documents in which the term '*computer*' occurs is $df_{computer}$. For Doc1,

$$idf_{computer} = log_2\left(N/df_{computer}\right)$$
$$= log_2(3/2)$$
$$= 0.5849$$

tf-idf weight for term '*computer*' is,
$$w_{computer} = tf_{computer} * idf_{computer}$$
$$= 60 * 0.5849$$
$$= 35.094$$

Similarly,
$$idf_{JAVA} = log_2\left(N/df_{JAVA}\right)$$
$$= log_2(3/1)$$
$$= 1.5849$$
$$w_{JAVA} = tf_{JAVA} * idf_{JAVA}$$
$$= 30 * 1.5849$$
$$= 47.547$$

After *tf-idf* measure, more weight is given to '*JAVA*' (the distinguishing term) and the weight of '*computer*' is much less (since it appears in more documents), although their actual frequencies depict an entirely different picture in the vector of *Doc1* above. The vector in *tf-idf* formulation can then be normalized using (4) to obtain the unit vector of the document.

III. MEASURING SIMILARITY

The most important factor in a clustering algorithm is the similarity measure [8]. In order to find the similarity of two vectors, *Cosine similarity* is used. For cosine similarity, the two vectors are multiplied, assuming they are normalized [2]. For any two vectors *v1, v2* normalized using (4),

Cosine *Similarity* (*v1, v2*) =

< (a1, c1), (a2, c2)...> . <(x1, k1), (x2, k2), (x3, k3)...>
$$= (c1)(k1) + (c2)(k2) + (c3)(k3) + ...$$

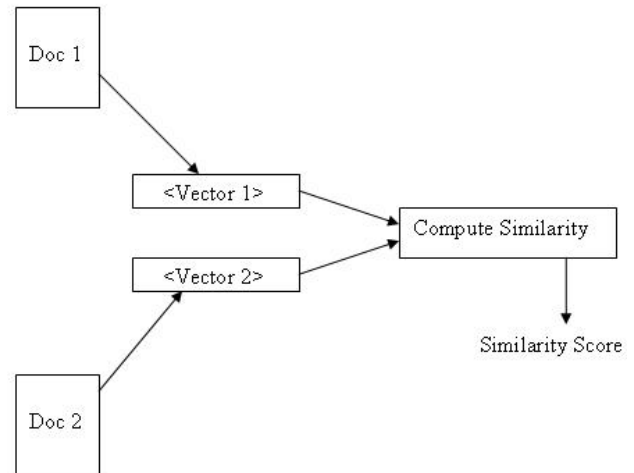where '.' is the ordinary dot product (a scalar value).

Figure 2. Computing Similarity

IV.   REPRESENTING A CLUSTER

The cluster is represented by taking average of all the constituent document vectors in the cluster. This results in a new summarized vector. This vector, like other vectors can be compared with other vectors, therefore, comparison between *document-document* and *document-cluster* follows the same method discussed in section III.

For cluster '*c*' containing two documents,
- ■ *v1 < (a, p1), (b, p2)...>*
- ■ *v2 < (a, q1), (b, q2)...>*

cluster representation is merely a matter of taking vector average of the constituent vectors and representing it as a *composite document* [2]. i.e. a vector as the average (or mean) of constituent vectors
   Cluster {v1, v2} = < (p1+q1)/2, (p2+q2)/2...>
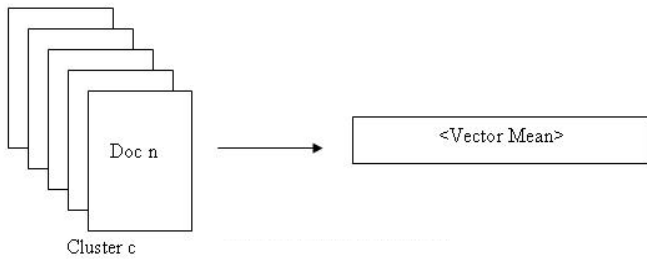
Figure 3.   Cluster Representation

V.   CLUSTERING ALGORITHM

The algorithm that we implemented is *k*-Means clustering algorithm. This algorithm takes *k*, number of initial bins as parameter and performs clustering. The algorithm is provided below [2]:

TABLE IV.   THE *K*-MEANS CLUSTERING ALGORITHM

1.   Distribute all documents among k bins.
     A bin is an initial set of documents that is used before the algorithm starts. It can also be considered as initial cluster.
 	
   a.   *The mean vector of the vectors of all documents is computed and is referred to as 'global vector'.*
   b.   *The similarity of each document with the global vector is computed.*
   c.   *The documents are sorted on the basis of similarity computed in part b.*
   d.   *The documents are evenly distributed to k bins.*
2.   Compute mean vector for each bin.
     As discussed in section IV.
3.   Compare the vector of each document to the bin means and note the mean vector that is most similar.
     As discussed in section III.
4.   Move all documents to their most similar bins.
5.   If no document has been moved to a new bin, then stop; else go to step 2.

VI.   DETERMINING *K*, NUMBER OF CLUSTERS

k-Means algorithm takes *k*, number of bins as input, therefore the value of *k* cannot be determined in advance without analyzing the documents. *k* can be determined by first performing clustering for all possible cluster size and then selecting the *k* that gives the minimum total variance, *E(k)* (error) of documents with their respective clusters. Note that the value of *k* in our case ranges from 2 to N. Clustering with *k=1* is not desired as single cluster will be of no use. For all the values of *k* in the given range, clustering is performed and variance of each result is computed as follows [2]:

$$E(k) = \sum_{i=1}^{n} \frac{(x^i - m_{ci})^2}{n}$$

where x^i is the *i-th* document vector, m_{ci} is its cluster mean and $c_i \in \{1,....., k\}$ is its corresponding cluster index.

Once the value of *k* is determined, each cluster can be assigned a label by using categorization algorithm [2].

VII. CLUSTERING RESULT

An input sample of 24 documents [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34] were provided to the *k*-Means Algorithm. With the initial value of *k*=24, the algorithm was run for three different scenarios:

(a) When the document vectors were formed on the basis of features (words) of the document.

(b) When the document vectors were formed on the basis of sub-category of features.

(c) When the document vectors were formed on the basis of parent category of the feature.

The result of *k*-means clustering algorithm for each case is given below:

TABLE V.   CLUSTERS- ON THE BASIS OF FEATURE VECTORS

Cluster name	Documents
Text Mining	[12, 13, 14, 16, 17, 18, 28]
Databases	[11, 25, 26, 27]
Operating Systems	[23, 32]
Mobile Computing	[22, 24]
Microprocessors	[33, 34]
Programming	[30, 31]
Data Structures	[29]
Business Computing	[20, 21]
World Wide Web	[15]
Data Transfer	[19]

TABLE VI.   CLUSTERS – ON THE BASIS OF SUBCATEGORY VECTORS[4]

Cluster name	Documents
Text Mining	[12, 13, 14, 16, 17, 18, 28, 31]
Databases	[11, 25, 26, 27]
Operating Systems	[21, 23, 32]
Communication	[22, 24]
Microprocessors	[33, 34]
Programming Languages	[30]
Data Structures	[29]
Hardware	[20]
World Wide Web	[15]
Data Transfer	[19]

TABLE VII.   CLUSTERS – ON THE BASIS OF PARENT CATEGORY VECTORS[4]

Cluster name	Documents
Software	[11, 12, 14, 16, 17, 25, 26, 27, 28, 30]
Operating Systems	[22, 23, 24]
Hardware	[31, 32, 33, 34]
Text Mining	[13, 18]
Network	[19, 20, 21 29]
World Wide Web	[15]

[4]the decision of selecting parent category vectors or sub-category vectors depends on the total number of root (parent) level categories, levels of sub-categories and organization of the domain dictionary used. A better, rich and well organized domain dictionary directly affects document representation; yields better clustering result and produces more relevant cluster names.
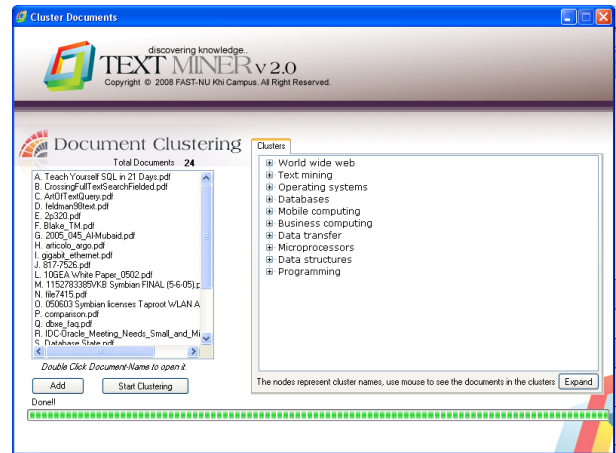
Figure 4.   Document Clustering in our Text Mining Tool

VIII. TECHNIQUE FOR IMPROVING THE QUALITY OF CLUSTERS

A.   Using Domain Dictionary to form vectors on the basis of sub-category and parent category

The quality of clusters can be improved by utilizing the domain dictionary which contains words in a hierarchical fashion.
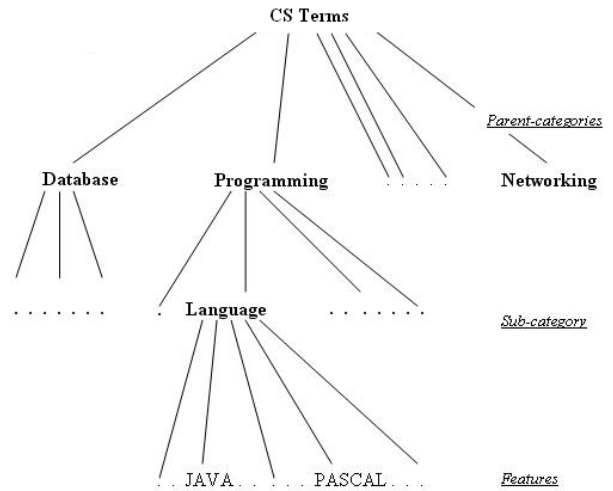
Figure 5.   Domain Dictionary (CS Domain)

For every word w and sub-category s,

$$w R s \qquad (6)$$

iff w comes under the sub-category s in the domain dictionary, where R is a binary relation.

The sub-category vector representation of a document with features,

$< (w_1, f_{w1}), (w_2, f_{w2})... (w_n, f_{wn}) >$
is
$< (s_1, f_{s1}), (s_2, f_{s2})... (s_m, f_{sm})>$
where *n* is the total number of unique features (words) in the document.

$$\forall n \; \exists m \ni w_n \; R \; s_m$$

for some $1 <= m <= c$ (c is total number of sub-categories)
f_{wn} is the frequency of the *n-th* word
f_{sm} is the frequency of *m-th* sub-category
R is defined in (6)

Consider a *feature* vector with features (words in a document) as vector dimension:

Document1< (register, 400), (JAVA, 22)... >

The *sub- category* vector of the same document is:

Document1<(architecture, 400+K_1), (language, 22+K_2)..>

where K_1 and K_2 are the total frequencies of other features that come under the sub-category '*architecture*' and '*language*' respectively.

Sub-category and parent category vectors generalize the representation of document and the result of document similarity is improved.

Consider two documents that are written on the topic of '*Programming language*', both documents are similar in nature but the difference is that one document is written on programming JAVA and the other on programming PASCAL. If document vectors are made on the basis of *features*, both the documents will be considered less similar because not both the documents will have the term 'JAVA' or 'PASCAL' (even though both documents are similar as both come under the category of *programming* and should be grouped in same cluster).

If the same documents are represented on the basis of sub-category vectors then regardless of whether the term JAVA occurs or PASCAL, the vector dimension used for both the terms will be '*programming language*' because both 'JAVA' and 'PASCAL' come under the sub-category of '*programming language*' in the *domain dictionary*. The similarity of the two documents will be greater in this case which improves the quality of the clusters.

IX. FUTURE WORK

So far our work is based on predictive methods using frequencies and rules. The quality of result can be improved further by adding English Language semantics that contribute in the formation of vectors. This will require incorporating some NLP techniques such as POS tagging (using Hidden Markov Models, HMM) and then using the tagged terms to determine the importance of features. A tagger finds the most likely POS tag for a word in text. POS taggers report precision rates of 90% or higher [10]. POS tagging is often part of a higher-level application such as Information Extraction, a summarizer, or a Q&A system [1]. The importance of the feature will not only depend on the frequency itself, but also on the context where it is used in the text as determined by the POS tagger.

X. CONCLUSION

In this paper we have discussed the concept of document clustering. We have also presented the implementation of k-means clustering algorithm as implemented by us. We have compared three different ways of representing a document and suggested how an organized domain dictionary can be used to achieve better similarity results of the documents. The implementation discussed in this paper is limited only to predictive methods based on frequency of terms occurring in the document, however, the area of document clustering needs to be further explored using language semantics and context of terms. This could further improve similarity measure of documents which would ultimately provide better clusters for a given set of documents.

REFERENCES

[1] Manu Konchady, 2006, *"Text Mining Application Programming"*. Publisher: Charles River Media. ISBN-10: 1584504609.

[2] Scholom M.Weiss, Nitin Indurkhya, Tong Zhang and Fred J. Damerau, *"Text Mining, Predictive Methods for Analysing Unstructured Information"*. Publisher: Springer, ISBN-10: 0387954333.

[3] Cassiana Fagundes da Silva, Renata Vieira, Fernando Santos Osório and Paulo Quaresma, *"Mining Linguistically Interpreted Texts"*.

[4] Martin Rajman and Romaric Besancon, *"Text Mining - Knowledge Extraction form Unstructured Textual Data"*.

[5] T. Nasukawa and T. Nagano, *"Text Analysis and Knowledge Mining System"*.

[6] Haralampos Karanikas, Christos Tjortjis and Babis Theodoulidis, *"An Approach to Text Mining using Information Extraction"*.

[7] Raymond J. Mooney and Un Yong Nahm, *"Text Mining with Information Extraction"*.

[8] Haralampos Karanikas and Babis Theodoulidis, *"Knowledge Discovery in Text and Text Mining Software"*.

[9] Alexander W. Dent, *"Fundamental problems in provable security and cryptography"*.

[10] Eric Brill, 1992 *"A Simple Rule-Based Part of Speech Tagger"*.

[11] *"Teach Yourself SQL in 21 Days"*, Second Edition, Publisher: MACMILLAN COMPUTER PUBLISHING USA.

[12] *"Crossing the Full-Text Search /Fielded Data Divide from a Development Perspective"*. Reprinted with permission of PC AI Online Magazine V. 16 #5.

[13] *"The Art of the Text Query"*. Reprinted with permission of PC AI Online Magazine V. 14 #1.

[14] Ronen Feldman1, Moshe Fresko1, Yakkov Kinar et al., *"Text Mining at the Term Level"*.

[15] Tomoyuki Nanno, Toshiaki Fujiki et al., *"Automatically Collecting, Monitoring, and Mining Japanese Weblogs"*.

[16] Catherine Blake and Wanda Pratt, *"Better Rules, Fewer Features: A Semantic Approach to Selecting Features from Text"*.

[17] Hisham Al-Mubaid, *"A Text-Mining Technique for Literature Profiling and Information Extraction from Biomedical Literature"*, NASA/UHCL/UH-ISSO. 49.

[18] L. Dini and G. Mazzini, *"Opinion classification through Information Extraction"*.

[19] Intel Corporation, *"Gigabit Ethernet Technology and Solutions"*, 1101/OC/LW/PP/5K NP2038.

[20] Jim Eggers and Steve Hodnett, *"Ethernet Autonegotiation Best Practices"*, Sun BluePrints™ OnLine—July 2004.

[21] 10 Gigabit Ethernet Alliance, *"10 Gigabit Ethernet Technology Overview"*, Whitepaper Revision 2, Draft A • April 2002.

[22] Catriona Harris, *"VKB Joins Symbian Platinum Program to Bring Virtual Keyboard Technology to Symbian OS Advanced Phones"*, for immediate release, Menlo Park, Calif. – May 10, 2005.

[23] Martin de Jode, March 2004, *"Symbian on Java"*.

[24] Anatolie Papas, *"Symbian to License WLAN Software from TapRoot Systems for Future Releases of the OS Platform"*, for immediate release LONDON, UK – June 6th, 2005.

[25] David Litchfield, November 2006, *"Which database is more secure? Oracle vs. Microsoft"*. Publisher: An NGSSoftware Insight Security Research (NISR).

[26] Oracle, 2006, *"Oracle Database 10g Express Edition FAQ"*.

[27] Carl W. Olofson, 2005, *"Oracle Database 10g Standard Edition One: Meeting the Needs of Small and Medium-Sized Businesses"*, IDC, #05C4370.

[28] Ross Anderson, Ian Brown et al., *"Database State"*. Publisher: Joseph Rowntree Reform Trust Ltd., ISBN 978-0-9548902-4-7.

[29] Chris Okasaki, 1996, *"Purely Functional Data Structures"*. A research sponsored by the Advanced Research Projects Agency (ARPA) under Contract No. F19628-95-C-0050.

[30] Jiri Soukup, *"Intrusive Data Structures"*.

[31] Anthony Cozzie, Frank Stratton, Hui Xue, and Samuel T. King, *"Digging for Data Structures"*, 8th USENIX Symposium on Operating Systems Design and Implementation pp. 255-266.

[32] Jonathan Cohen and Michael Garland, *"Solving Computational Problems with GPU Computing"*, September/October 2009, Computing in Science & Engineering.

[33] R. E. Kessler, E. J. McLellan1, and D. A. Webb, *"The Alpha 21264 Microprocessor Architecture"*.

[34] Michael J. Flynn, *"Basic Issues in Microprocessor Architecture"*.

AUTHORS PROFILE

Yasir Safeer received his BS degree in Computer Science from FAST - National University of Computer and Emerging Sciences, Karachi, Pakistan in 2008. He was also awarded Gold Medal for securing 1st position in BS in addition to various merit based scholarships during college and undergraduate studies. He is currently working as a Software Engineer in a software house. His research interests include text mining & information extraction and knowledge discovery.

Atika Mustafa received her MS and BS degrees in Computer Science from University of Saarland, Saarbruecken, Germany in 2002 and University of Karachi, Pakistan in 1996 respectively. She is currently an Assistant Professor in the Department of Computer Science, National University of Computer and Emerging Sciences, Karachi, Pakistan. Her research interests include text mining & information extraction, computer graphics(rendering of natural phenomena, visual perception).

Anis Noor Ali received his BS degree in Computer Science from FAST - National University of Computer and Emerging Sciences, Karachi, Pakistan in 2008. He is currently working in an IT company as a Senior Software Engineer. His research interests include algorithms and network security.

Controlling Wheelchair Using Electroencephalogram

Vijay Khare[1]

Dept. of Electronics and Communication, Engineering

Jaypee Institute of Information Technology
Nioda, India
Email : vijay.khare@jiit.ac.in

Sneh Anand[3]
Centre for Biomedical Engineering Centre
Indian Institute of Technology,
Delhi, India
Email : sneh@iitd.ernet.in

Jayashree Santhosh[2]
Computer ServicesCentre

Indian Institute of Technology,
Delhi, India
 Email : jayashree@cc.iitd.ac.in

Manvir Bhatia[4]
 Department of Sleep Medicine,
 Sir Ganga Ram Hospital,
New Delhi, India
 Email : manvirbhatia1@yahoo.com

Abstract— **This paper present the development of a power wheelchair controller based on Electroencephalogram (EEG).To achieve this goal   wavelet packet transform (WPT) was used for feature extraction of the relevant frequency bands from electroencephalogram (EEG) signals. Radial Basis Function network was used to classify the pre defined movements such as rest, forward, backward, left and right of the wheelchair. Classification and evaluation results showed the feasibility of EEG as an input interface to control a mechanical device like powered wheelchair.**

  Keywords— Electroencephalogram (EEG), Wavelet Packet Transform (WPT), Radial Basis Function neural network (RBFNN), Brain computer interface (BCI), Rehabilitation, Wheelchair Controller.

I.   INTRODUCTION

There are numerous interfaces and communication methods between human and machines. A typical human interface utilizes input devices such as keyboard, mouse, joystick, chin control, ultrasonic non contact head controller and voice controller. Such interfaces were developed to improve manipulability, safety and comfortness. Literature survey shows existing systems such as   Chin controller is inconvenient to use, ultrasonic non-contact head controller has relatively low accuracy and voice controller gives delayed response to voice command hence not useful in noisy environment[1-2]. Recently, a number of biological signals such as electromyogram (EMG), Electroencephalogram (EEG) and Electrocculogram (E.O.G) have been employed as hands-free interface to machines [3-7]. Brain Computer Interface (BCI) system has been shown to have the potential to offer humans a new nonmuscular communication channel, which enables the user to communicate with their external surroundings using the brain's electrical activity measured as electroencephalogram (EEG) [8-12].

This paper introduces the working prototype of a Brain Controlled Wheelchair (BCW) that can navigate inside a typical office and hospital environment with minimum structural modification. It is safe and relatively low cost and provides optimal interaction between the user and wheelchair within the constraints of brain computer interface.

In this study, Wavelet Packet Transform (WPT) method was used for feature extraction of mental tasks from eight channel EEG signals. WPT coefficients give the best discrimination between the directions of wheelchair in the relevant frequency band. The WPT coefficients were used as the best fitting input vector for classifier. Radial Basis Function network was used to classify the signals.

II.   METHODOLOGY

A.   Subjects

Nine right-handed healthy male subjects of age (mean: 23yr) having no sign of any motor- neuron diseases were selected for the study. A pro-forma was filled in with detail of their age & education level as shown in Table I. The participants were student volunteers for their availability and interest in the study. EEG data was collected after taking written consent for participation. Full explanation of the experiment was provided to each of the participants.

TABLE I.   CLINICAL CHARACTERISTICS OF SUBJECTS

S.No.	Subject	Age	Educational status
1	Subject 1	22	BE
2	Subject 2	21	BE

3	Subject 3	23	BE
4	Subject 4	27	M.TECH
5	Subject 5	23	BE
6	Subject 6	22	BE
7	Subject 7	27	M.TECH
8	Subject 8	22	BE
9	Subject 9	22	BE

B. EEG Data Acquisition

EEG Data used in this study was recorded on a Grass Telefactor EEG Twin3 Machine available at Deptt. of Neurology , Sir Ganga Ram Hospital, New Delhi. EEG recording for nine selected subjects were done for five mental tasks for five days. Data was recorded for 10 sec during each task and each task was repeated five times per session per day. Bipolar and Referential EEG was recorded using eight standard positions C3, C4, P3, P4, O1 O2, and F3, F4 by placing gold electrodes on scalp, as per the international standard 10-20 system of electrode placement as shown in Fig 1. The reference electrodes were placed on ear lobes and ground electrode on forehead. EOG (Electooculargram) being a noise artifact, was derived from two electrodes placed on outer canthus of left and right eye in order to detect and eliminate eye movement artifact. The settings used for data collection were: low pass filter 1Hz, high pass filter 35 Hz, sensitivity 150 micro volts/mm and sampling frequency fixed at 400 Hz.
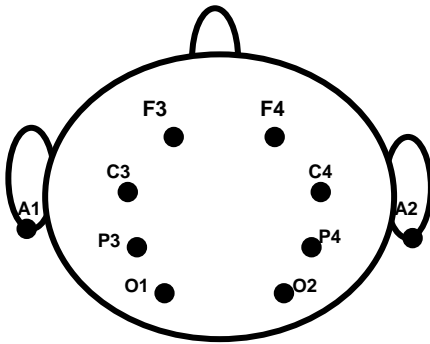

Figure1:- Montage for present study

C. Experiment Paradigm

An experiment paradigm was designed for the study and the protocol was explained to each participant before conducting the experiment. In this, the subject was asked to comfortably lie down in a relaxed position with eyes closed. After assuring the normal relaxed state by checking the status of alpha waves, the EEG was recorded for 50 sec, collecting five session of 10sec epoch each for the relaxed state. This was used as the baseline reference for further analysis of mental task. The subject was asked to perform a mental task on presentation of an audio cue. Five session of 10sec epoch for each mental task were recorded, each with a time gap of 5

minute (as shown in Fig 2). The whole experiment lasted for about one hour including electrode placement.
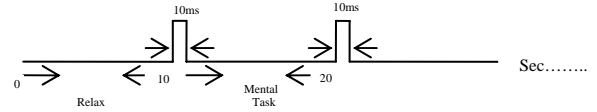
Figure 2: Timing of the Protocol

Data collected from nine subjects performing five mental tasks were analyzed. The following mental tasks were used to record the appropriate EEG data.

- Movement Imagination:-The subject was asked to plan movement of the right hand.

- Geometric Figure Rotation:-The subject was given 30 seconds to see a complex three dimensional object, after which the object was removed. The subject was instructed to visualize the object being rotated about an axis.

- Arithmetic Task:-The subject was asked to perform trivial and nontrivial multiplication. An example of a trivial calculation is to multiply 2 by 3 and nontrivial task is to multiply 49 by 78. The subject was instructed not to vocalize or make movements while solving the problem.

- Relaxed: - The subject was asked to relax with eyes closed. No mental or physical task to be performed at this stage.

D. Feature Extraction

The frequency spectrum of the signal was first analyzed through Fast Fourier Transform (FFT) method [13-14]. The FFT plots of signals from all the electrode pairs were observed and maximum average change in EEG amplitude was noted as shown in Fig3. For relaxed state, the peak of power spectrum almost coincides at for central and occipital area in the alpha frequency range (8-13Hz) [15]. EEG recorded with relaxed state is considered to be the base line for the subsequent analysis. Mu rhythms are generated over sensorimotor cortex during planning a movement. For movement imagery of right hand, maximum upto 50% band power attenuation was observed in contralateral (C3 w.r.t C4) hemisphere in the alpha frequency range (8-13Hz) [16]. For geometrical figure rotation, the peak of the power spectrum was increased in right hemisphere rather than left in the occipital area for the alpha frequency range (8-13Hz)[17]. For trivial multiplication, the peak of the power spectrum was increased in left hemisphere rather than right hemisphere in the frontal area for the alpha frequency range (8-13Hz)[18].For non trivial multiplication, the peak of the power spectrum was increased in left hemisphere rather than right

hemisphere in the parietal area for the alpha frequency range (8-13Hz).

Figure 3: Maximum Average change in Amplitude of PSD

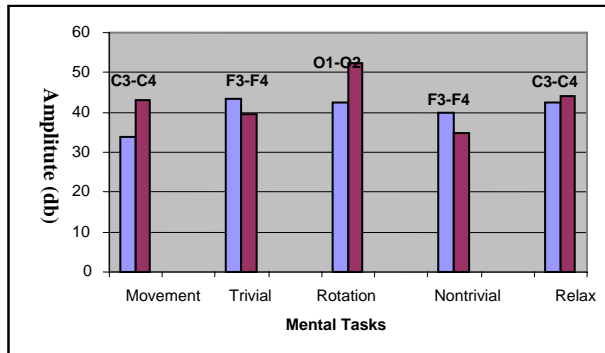The data was preprocessed using Wavelet packet transform to extract the most relevant information from the EEG signal. [19-20]. By applying Wavelet packet transform on the original signal wavelet coefficients in the (8-13Hz) frequency band at the 5th level node (5, 3) were obtained. Twenty one coefficients have been obtained from one second of EEG data. These coefficients are scaled and used as the best fitting input vector for classifiers. Subsequently the signal was reconstructed at node (5, 3).

E. Classifier

For classification, Radial Basis Function Neural Network (RBFNN) classifier was employed. A two layer network was implemented with 21 input vectors, a hidden layer with Gaussian activation function consisting as many as hidden neurons as input vectors and five neuron in the output layer [21-23]. RBFNN produces a network with zero error on training vectors. Using RBFNN the five mental tasks were classified, as shown in Tables II

TABLE II.   CLASSIFICATION OF FOUR MENTAL TASKS

Tasks	Accuracy %	classifications
Movement Imagery	100	00100
Trivial Multiplication	100	01000
Geometric Figure Rotation	100	00010
Nontrivial Multiplication	100	10000
Relax	100	00001

After the classification of five mental tasks namely movement imagery, trivial multiplication, geometrical figure rotation, nontrivial multiplication and relax, the output of the classifier was interfaced with the motor using parallel port. The motor driver required 3 bit of data. The output of classifier was

mapped into 3 bit as shown in table3 to provide parallel port input bit, which was used to drive the motor

F. Hardware implementation
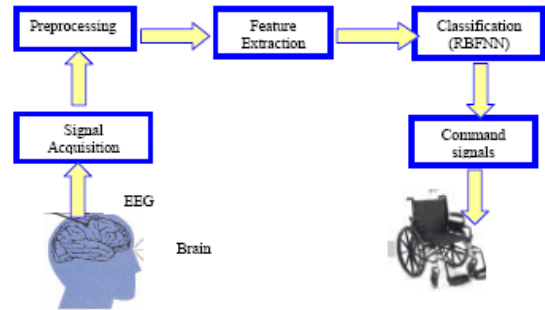
Figure 4: Conceptual block diagram of the wheelchair controlled by EEG signals (courtesy http//www. getch.at/g.bcisys/bci.html)

Conceptual block diagram of EEG based power wheelchair system is shown in Fig 4. Using parallel port, Motor driver IC (IC L293) was interfaced with computer as shown in Fig 5 for the wheelchair controller. In the circuit, P1 acts to enable the chip and combination of P2 and P3 were used to control direction of wheelchair. The truth table for the above logic is shown in Table III with polarities of motor of M1and M2. All five direction of wheelchair movement were properly controlled by this designed circuit.
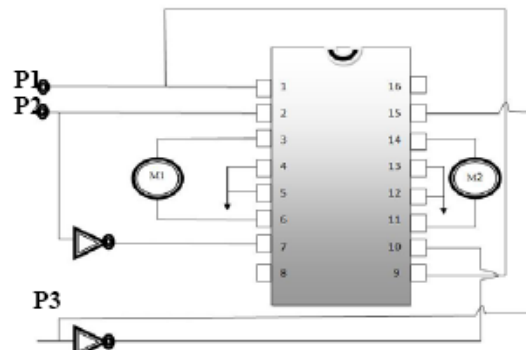
Figure 5: Circuit Diagram for wheelchair controller

TABLE III.   TRUTH TABLE OF HARDWARE DESIGN

P1	P2	P3	*M1* +	*M1* -	*M2* +	*M2* -	TASKS
1	0	0	0	1	1	0	LEFT (L)
1	1	0	1	0	1	0	FORWARD(F)
1	0	1	0	1	0	1	BACKWARD(B)
1	1	1	1	0	0	1	RIGHT (R)
0	X	X	---	---	---	---	STOP(S)
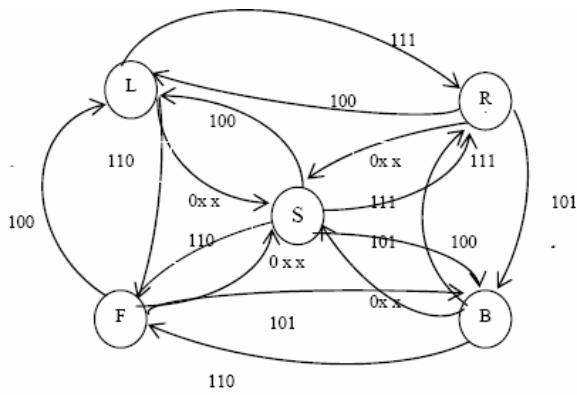
Figure 6: State diagram for Wheelchair Movement

The polarities of the motors M1 and M2 are shown in the truth table. State diagram for wheel chair movement in different direction is shown in Fig 6. For movement imagery task, the output of parallel port would be [1 0 0]. Due to opposite polarities, M2 motor would move forward and M1motor backward which would be lead to left movement of the wheelchair. For trivial multiplication task, the output of parallel port would be [1 1 0]. Due to same polarities, both motors M1and M2 move forward resulting forward movement of the wheelchair. For geometrical figure rotation task, the output of parallel port would be [1 1 1]. Due to opposite polarities, M1 motor moves forward and M2 motor backward resulting right movement of the wheelchair. For nontrivial multiplication, the output of parallel port would be [1 0 1]. Due to same polarities, both motors M1and M2 move backward resulting backward movement of the wheelchair. Similarly, for stop tasks output of parallel port would be [0 x x] and the wheelchair would be control by different polarities at the motors.

III. RESULT AND DISCUSSION

Classification of five mental tasks shown in the Fig (7-11) Earlier researchers had established [15-18] the most prominent areas in brain for domain of information during various mental tasks as shown in Table IV. In the present study, maximum average change in EEG amplitude has been observed by us as shown in TableV. The study had led to following observations:

- For movement imagery task, the amplitude of the power spectrum for alpha frequency range (8-13Hz) had attenuation in contralateral area.

- For geometrical figure rotation task, the amplitude of the power spectrum increases in the right occipital region for alpha frequency range (8-13Hz).

- For trivial multiplication task, the amplitude of the power spectrum increases in the left frontal region for alpha frequency range (8-13Hz).

- For nontrivial multiplication task, the amplitude of the power spectrum increases in the left parietal region for alpha frequency range (8-13Hz).

This observation could be used successfully for controlling power wheelchair. It can be noted that comparing Table IV&V, there is perfect match with earlier studies and changes are prominent and unique.
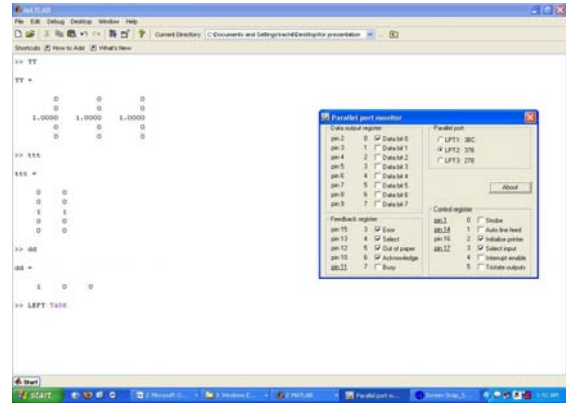
Figure 7: Movement task classification
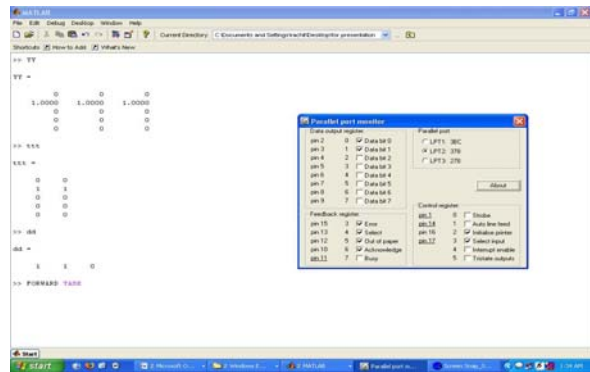
Figure8: Trivial multiplication task classification
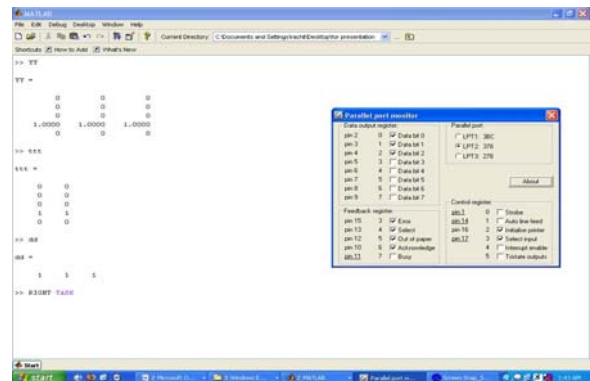
Figure9: Geometrical figure rotation classification

Figure10: Non trivial task classification

Figure11 Relax task classification

TABLE IV.   DOMAIN OF INFORMATION

Tasks	Domain of information	(Contralateral/ Ipsilateral)	Type of change in amplitude of alpha rhythm(8-13Hz)
Movement Imagination	Central	Contralateral	Decreased
Arithmetic Simple	Frontal,	Ipsilateral	Increased
Geometrical figure rotational	Occipital	Ipsilateral	Increased
Arithmetic complex	parietal	Ipsilateral	Increased
Base line	Occipital, Central	Contralateral	Coincide

TABLE V.   AVERAGED AMPLITUDE IN ALPHA FREQUENCY FOR NINE SUBJECT

Tasks	Electrodes		Amplitude(db) in alpha rhythm(8-13Hz)	
Movement Imagination	C3	C4	C3(33.75)	C4(43.02)
Trivial Multiplication	F3	F4	F3(43.34)	F4(39.69)
Geometrical figure rotational	O1	O2	O1(42.35)	O2(52.43)
Non-trivial Multiplication	P3	P4	P3(40.03)	P4(34.74)
Relax	C3	C4	C3(42.29)	C4(44.01)

Fig 12(a-d) associated with table 6 show four experiments conducted on nine right-handed male subjects. The subjects were asked to mentally drive the wheelchair from the starting point to a goal by executing the five different mental tasks namely Movement Imagery (MI), Trivial Multiplication(TM), Geometrical Figure Rotation (GFR), Non Trivial Multiplication (NTM) and Relax (R) to control direction of the power wheelchair.

To complete task from staring point to goal, the subject performed sequence of the mental tasks as shown in TableVI. Experiment has been successfully completed with 100% accuracy by all nine subjects.

TABLE VI.   MATRIX OF MENTAL TASKS AND DIRECTION OF WHEELCHAIR

Path a	TM/ Forward	GFR/ Left	GFR/Left	GFR/ Left	R/ Stop
Path b	TM/ Forward	MI/ Right	MI/ Right	MI/ Right	R/ Stop
Path c	TM/ Forward	GFR/ Left	GFR/Left	GFR/ Left	R/ Stop
Path d	TM/ Forward	MI/ Right	GFR/Left	GFR/ Left	R/ Stop

Figure12(a-d): Top view of random path

IV. CONCLUSION

The present study investigated controlling of a power wheelchair by EEG signals. This was an attempt to control direction of wheelchair via brain signals. Each direction (left, forward, right, backward and stop) of the wheelchair corresponded to five mental tasks (movement imagery, trivial multiplication, geometrical figure rotation, non -trivial multiplication and relax). To differentiate five mental tasks, wavelet packet transform was employed for feature extraction and Radial basis function neural network was used for classification. The experimental result showed 100% accuracy.

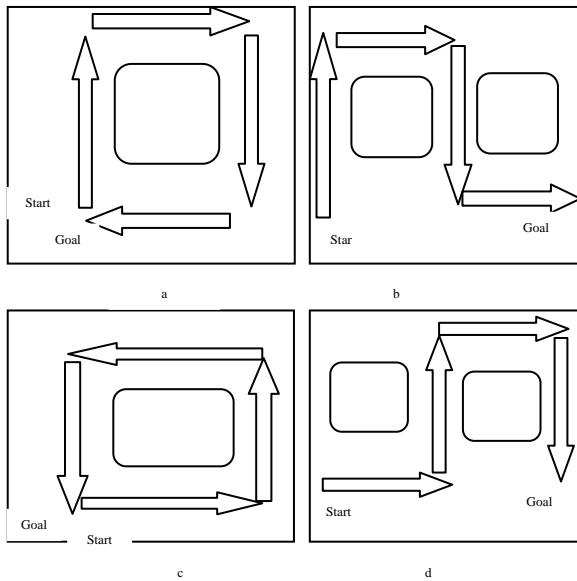The authors would like to extend the work with severely disabled people and to customize the device as per individual response and requirements. This kind of system can also be used in a variety of applications like–

➢ Environment control units (ECU'S)
➢ Helping disable people to directly interact with hand held devices such as cell phones and PDAs.
➢ Dealing with hazardous material/chemical at laboratories.

ACKNOWLEDGMENT

REFERENCES

[1] K. Tanaka. K. Matsunaga, N. Kanamori, S. Hori, and H.O.Wang, "Electroencephalogram based control of a mobile robot," in proc. IEEE Int. Symp. Computational Intell.. Robot.Autom Kobe, Japan, pp 670-675, Jul.2003.

[2] K. choi & A. Cichocki , "Control a Wheelchair by Motor Imagery in Real Time," IDEAL 2008, LNCS 5326 , pp. 330-337, Springer Verlag Berlin Heidelberg 2008.

[3] F. Galan, M. Nuttan, E. Lew, P.W. Ferrez, G. Vanacker, J. Philip, J. del R.Millan, "A brain actuated wheelchair: Asynchronous and nojn invasive brain computer interfaces for continuous control of robot" ,Clinical Neurophysiology, Vol 119, pp. 2159-2169, 2008.

[4] A.T.C. Au and R.F. Kirsch, "EMG based prediction of shoulder and elbow kinematics in able –bodies and spinal cord injured individual," IEEE Trans. Rehab.Eng.vol.8 no.4, pp. 471-480, Dec.2000.

[5] J. Millan, "Noninvasive brain actuated control of a mobile robot by human EEG," IEEE Trans. Biomed. Eng, vol. 51, no 6, pp.1026-1033, June 2004.

[6] R. Bare, "E.O.G guidance of a wheeelchair using neural network," in porc.Int.Conf. Pattern recognition, Barcelona Spain, pp. 4668-4672, 2000.

[7] R. Leeb, "Self paced (Asynchronous) BCI-Contol of a wheelchair in virtual environment :Acase study with tetraplegic", Computational Intelligent & Neuroscience.Vol 2007, Aritical ID 79642, 2007

[8] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, B Arnaldi, "A Review of Classification algorithms for EEG bases brain computer interface," Journal of neural Engineering, Vol. 4, R1-R13, 2007.

[9] R. Boostani, B. Graimann, M.H. Moradi, G. Plurfscheller, " Comparison approach toward finding the best feature and classifier in cue BCI", Med. Bio. Engg., Computer Vol 45, pp. 403-413, 2007.

[10] J.R. Wolpaw, N. Birbaumer, D.J. Mc Farland, G.Plurtscheller, T.M. Vaughan, "Brain computer Interfaces for communication and control", Clinical Neurophys. , Vol 113, pp. 767-791, 2002.

[11] G. Pfurtschelle, D. Flotzinger, and J.Kalcher, "Brain Computer interface-A new communication device for handicapped people," J Microcomput. Applicate., vol. 16, pp. 293-299, 1993.

[12] J. R. Wolpaw, T. M. Vaughan and E. Donchin, "EEG Based Communication prospects and problems," IEEE, Trans.Rehab. Eng., Vol.4, pp. 425-430, Dec.1996.

[13] Z. A. Keirn and J. I. Aunon, "A new mode of communication between man and his surroundings," IEEE Trans. Biomed. Eng., Vol. 37, No. 12, pp. 1209–1214, Dec. 1990.

[14] R. Palaniappan, "Brain computer interface design using band powers extracted during mental task," proceeding of the 2nd International IEEE EMBS Conference on Neural Engineering , pp. 321- 324, 2005.

[15] G. Pfurtscheller, C. Neuper, A. Schlogl and K. Lugger, "Separability of EEG signals recorded during right and left motor imagery using adaptive auto regressive parameters, IEEE. Trans. on rehabilitation Engineering, Vol 6 ,No3, pp. 316-325, 1998.

[16] J. Santhosh, M. Bhatia, S. Sahu, S. Anand, "Quantitative EEG analysis for assessment to plan a task in ALS patients, a study of executive function (planning) in ALS," Cognitive brain research Vol 22, pp. 59-66, 2004.

[17] A. R. Nikolaev and A. P. Anokhin, "EEG frequency ranges during reception and mental rotation of two and three dimensional objects," Neuroscience and Bheaviour physiology, Vol. 28, No-6, 1998.

[18] Osaka M., "Peak alpha frequency of EEG during a mental task: task difficulty and hemisphere difference," Psychophysiology, Vol.21, pp. 101-105, 1984.

[19] C. S. Li and H. Wang, "Wavelet transform for on –off switching BCI device," 7th Asian-Pacific Conference on Medical and Biological Engineering, Beijing, China, Vol 19, pp. 363-365, 22–25 April 2008.

[20] B. Guo Xu & A. G. Song , "Pattern recognition of motor imagery EEG using wavelet transform," Jouranl of Biomedical Science &Engineering, Vol 1, pp. 64-67, 2008.

[21] E. Larsson, K. Åhlander and A. Hall, "Multi-dimensional option pricing using radial basis functions and the generalized Fourier transform," In J. Comput. Appl. Math., 2008.

[22] U. Pettersson, E. Larsson, G. Marcusson and J. Persson, "Improved radial basis function methods for multi-dimensional option," In J. Comput. Appl. Math., 2008.

[23] S. Chen, C. F. N. Cowan and P. M. Grant, "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks," IEEE Transactions on Neural Networks, vol. 2, No. 2, pp. 302-309, March 1991.

AUTHORS PROFILE

Vijay Khare is currently pursuing his PhD in Bio Signal Processing at the Indian Institute of Technology, Delhi. He did his M.Tech in Instrumentation & Control, from NSIT Delhi. He is currently, with the Dept. Electronics and Communications Engineering at the Jaypee Institute of Information Technology. His research interests are Neural Networks, Brain Computer Interfacing, and Control Systems.

Dr.Jayashree Santhosh completed her B.Tech in Electrical Engineering from University of Kerala, M Tech in Computer & Information Sciences from Cochin University of Science and Technology, Kerala and Ph.D from IIT Delhi. She is a Fellow member of IETE, Life member of Indian Association of Medical Informatics (IAMI) and Indian Society of Biomechanics (ISB). Her research interests include IT in Healthcare Systems and was associated with a project on IT in Health Care at City University of Hong Kong. She is also associated with various projects with Centre for Bio-Medical Engineering at IIT Delhi in the area of Technology in Healthcare. Her research interests focus on Brain Computer Interface Systems for the Handicapped and in Neuroscience.

Prof. Sneh Anand is a professor and head, Center for Biomedical Engineering, Indian Institute of Technology, Delhi. She did B.Tech in Electrical Engg, from Punjab University, Patiala, and M.Tech in Instrumentation & Control from IIT Delhi and Ph.D. in Biomedical Engg. from IIT Delhi. Her research interests include biomedical instrumentation, rehabilitation engineering, biomedical transducers and Sensors.

Dr. Manvir Bhatia is the Chairperson of Dept. of SleepMedicine at Sir Ganga Ram Hospital, New Delhi and is also a Senior Consultant Neurologist.Dr. Manvir Bhatia completed her MBBS in 1981, and Doctor of Medicine in 1986 from Christian Medical College and Hospital, Ludhiana. DM in Neurology 1993, from All India Institute of Medical Sciences**.**She is a member of Indian Academy of Neurology, Indian Epilepsy Society, Indian Sleep Disorders Association, World Association of Sleep Medicine, International Restless Legs Society Study Group and American Academy of Electrodiagnostic Medicine.Dr. Manvir Bhatia has been invited to deliver lectures in National & International workshops, conferences on topics related to Neurology, Epilepsy, Sleep Medicine and has sleep published papers in leading journals.

A New Biometrics based Key Exchange and Deniable Authentication Protocol

K. Saraswathi
Asst.Proffessor, Department of Computer Science
Govt Arts College
Udumalpet, Tirupur, India
dharundharsan@rediffmail.com

Dr. R. Balasubramanian
Dean Academic Affairs
PPG Institute of Technology
Coimbatore, India
ramamurthybala2@gmail.com

Abstract-*Wireless Local Area Networks (WLANs) are gaining recognition as they are fast, cost effective, supple and easy to use. The networks face a serious of issues and challenges in establishing security to the users of the network. With users accessing networks remotely, transmitting data by means of the Internet and carrying around laptops containing sensitive data, ensuring security is an increasingly multifarious challenge. Therefore it is necessary to make sure the security of the network users. In order to provide network security many techniques and systems have been proposed earlier in literature. Most of these traditional methods make use of password, smart cards and so on to provide security to the network users. Though these traditional methods are effective in ensuring security they posses some limitations too. The problem with these traditional approaches is that there is possibility to forget the password. Moreover, compromised password lead to a fact, that unauthorized user can have access to the accounts of the valid user. This paper proposes an approach for network security using biometrics and deniable authentication protocol. The human biometrics like hand geometry, face, fingerprint, retina, iris, DNA, signature and voice can be effectively used to ensure the network security. The diverse phases included in this proposed approach are user registration, fingerprint enhancement, minutiae point extraction, mapping function and deniable authentication protocol. Furthermore, biometric authentication systems can be more convenient for the users since it involves no password that might be feared to be forgotten by the network users or key to be lost and therefore a single biometric trait (e.g., fingerprint) can be used to access several accounts without the burden of remembering passwords. This proposed paper also explains some of the fingerprint enhancement techniques to make the biometric template noise free. Experiments are conducted to evaluate the performance measure of the proposed approach.*

Keywords-*Biometrics, Cryptography, Data Security, Fingerprint, Mapping Function, Minutiae Point, Network Security, User Registration.*

I. INTRODUCTION

Accurate, automatic identification and authentication of users is an elemental problem in network environments. Shared secrets such as personal identification numbers or passwords and key devices like smart cards are not just enough in some cases. This authentication method has traditionally been based on passwords. The problem with these traditional approaches is that there is possibility to forget the password. Moreover, compromised password lead to a fact, that unauthorized user can have access to the accounts of the valid user. The Biometric based user authentication systems are highly secured and efficient to use and place total trust on the authentication server where biometric verification data are stored in a central database [1]. This biometrics based user

authentication system improves the network security. Some of most widely used biometric are hand geometry, face, fingerprint, retina, iris, DNA, signature and voice.

Biometrics is the science of measuring and statistically analyzing biological data can be used to recognize different body parts like the eyes, fingerprints, facial characteristics, voice etc. Thus, it takes security to the next level by not just confining it to authenticating passwords, fingerprint matching techniques [2]. Based on the individual's biometric characteristics a biometric system recognizes an individual. The process of a biometric system can be described, in a beginner's manner, by a three-step process. The foremost step in this process is collection of the biometric data which is formally known as user registration. This step uses different sensors, to assist the user in the registration process. The second step converts and describes the observed data using a digital representation called a template. This step varies between modalities and also between vendors. In the third step, the newly acquired template is compared with one or more previously generated templates stored in a database. The result of this comparison is a "match" or a "non-match" and is used for actions such as permitting access, sounding an alarm, etc [15].

Declaring a match or non-match is based on the obtained template being analogous, but not one and the same, to the stored template. A threshold determines the measure of similarity necessary to result in a match declaration. The acceptance or rejection of biometric data is completely dependent on the match score falling above or below the threshold. The threshold is adjustable so that the biometric system can be more or less stringent, depending on the requirements of any given biometric application [15]. Among all the biometric techniques, today fingerprints are the most widely used biometric features for personal identification because of their high acceptability, Immutability and individuality.

This paper proposes a technique to secure the network communication using biometric characteristics obtained from the individuals. The biometric characteristic used in this paper is fingerprint. This proposed paper utilizes image processing technique to extract the biometric measurement called minutiae from the user's fingerprint. The user's full finger print image is converted and stored as encrypted binary template, which is used for authentication by the server of the network. The user's biometric verification data are first transformed into a strong secret and is then stored in the

server's database during registration. The proposed system is evaluated to determine the performance measures.

The remainder of this paper is organized as follows. Section 2 discusses some of the related work proposed earlier in association to biometric based network security. Section 3 describes the proposed approach of providing network security using the biometric characteristics obtained fingerprint. Section 4 illustrates the performance measures and Section 5 concludes the paper with directions to future work.

II.   RELATED WORK

A lot of research has been carried out in the field of establishing network security based on biometric features obtained from individual user [13] [14]. This section of the paper discusses some of the related work proposed earlier in association to biometric based network security.

In their work [3] Rahman et al. proposed architecture for secure access of computers inside an organization from a remote location. They used biometrics features and a one-time password mechanism on top of secure socket layer (SSL) for authentication. Moreover they also provided three layers of security levels for network communication, and also a mechanism for secure file accesses based on the security privileges assigned to various users was proposed. The files to be accessed from the server are categorized depending on their access privileges and encrypted using a key assigned to each category. The test results of their approach evaluated the performance of their proposed approach.

Chung et al. in [4] described a method for biometric based secret key generation for protection mechanism. The binding of the user's identity and biometric feature data to an entity is provided by an authority through a digitally signed data structure called a biometric certificate. Therefore, the main goal (or contribution) of their work is to propose a simple method for generating biometric digital key with biometric certificate on fuzzy fingerprint vault mechanism. Biometric digital key from biometric data has many applications such as automatic identification, user authentication with message encryption, etc. Therefore, their work analyzed the related existing scheme and proposed a simplified model where a general fuzzy fingerprint vault using biometric certificate with security consideration.

Dutta et al. in [5] presented a novel method for providing network security using biometric and cryptography. They proposed a biometrics-based (fingerprint) Encryption/Decryption Scheme, in which unique key is generated using partial portion of combined sender's and receiver's fingerprints. From this unique key a random sequence is generated, which is used as an asymmetric key for both Encryption and Decryption. Above unique Key is send by the sender after watermarking it in sender's fingerprint along with Encrypted Message. The computational requirement and network security features are addressed. Proposed system has a advantage that for public key, it has not to search from a database and security is maintained.

Network security issues are projected by Benavente et al. in [6]. The Internet is increasingly becoming a public vehicle for remote operations. Integrating biometric information in the authentication chain explores new problems. Remote virtual identity is starting to play in the way towards an e-Europe, and applications for e-government integrate biometrics. Remote identity of subjects should be unambiguously stated. Several features drive the spread of biometric authentication in network applications, in order to provide end-to-end security across the authentication chain aliveness detection and fake-resistive methods, network protocols, security infrastructure, integration of biometrics and public key infrastructure (PKI), etc. Their paper proposed a mid-layer interoperable architecture furnished with a set of generic interfaces and protocol definitions. Their scheme enables a future introduction of new modules and applications with a minimal development effort.

An intelligent fingerprint based security system was designed and developed by Suriza et al. in [7]. Traditionally, user authentication is meant to provide an identification number or a password that is unique and well protected to assure the overall system security. This type of security system is very fragile in an area where a higher level of security system is required. Biometrics-based system offers a new and better approach to user authentication. Biometrics authentication is an automated method whereby an individual identity is confirmed by examining a unique physiological trait or behavioral characteristic, such as fingerprint, iris, or signature, since physiological traits have stable physical characteristics. The design and development of a fingerprint-based security system, comprising the scanner, interface system, Boltzmann machine neural network and access control system is discussed in this paper. The integration between the hardware and the software is completed by using Visual Basic 6 programming language. The results obtained both for the simulation studies and testing of the integrated system with real-life physical system have demonstrated the practicality of such system as well as its potential applications in many fields.

Ronald in [8] put forth an alternative approach for password in network security using biometrics. Passwords are the primary means of authenticating network users. However, network administrators are becoming concerned about the limited security provided by password authentication. Many administrators are now concluding that their password-based security systems are not all that secure. User passwords are routinely stolen, forgotten, shared, or intercepted by hackers. Another serious problem is that computer users have become too trusting. They routinely use the same password to enter both secure and insecure Web sites as well as their networks at work. In response to the proven lack of security provided by password authentication, network administrators are replacing network passwords with smartcards, biometric authentication, or a combination of the three. Smart cards are credit card-size devices that generate random numbers about every minute, in sync with counterparts on each entry point in the network. Smart cards work well as long as the card isn't stolen. A better

choice to ensure network security is the use of biometrics. Their paper investigated the different biometric techniques available to determine a person's identity. Also described, were the criteria for selecting a biometric security solution. In conclusion, efforts to establish biometric industry standards (including standard application program interfaces (APIs)) were discussed.

III. PROPOSED APPROACH

Biometric cryptosystems [9] join together cryptography and biometrics to promote from the strengths of both fields. In such systems, while cryptography provides high and adjustable security levels, biometrics brings in non-repudiation and eliminates the must to remember passwords or to carry tokens etc. In biometric cryptosystems, a cryptographic key is generated from the biometric template of a user stored in the database in such a way that the key cannot be revealed without a successful biometric authentication.

The overall architecture of the biometric system to improve network security is shown in figure 1. The Server maintains a database where the encrypted minutia template of the user's finger print is stored. In this setting, users communicate with the server for the principle of user authentication, by rendering users fingerprint, which is transformed into a long secret detained by the server in its database [1].
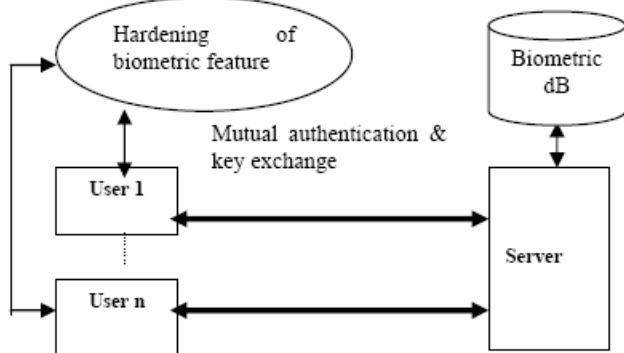

Figure 1.Biometric System

Figure 2 shows a common idea of obtaining the minutiae points from biometric feature obtained from the user. The key vector is formed based on minutiae points (ridge ending and ridge bifurcation) are encountered in the given finger print image [10]. Figure 2 shows various steps involved in the proposed system for network security using biometrics.
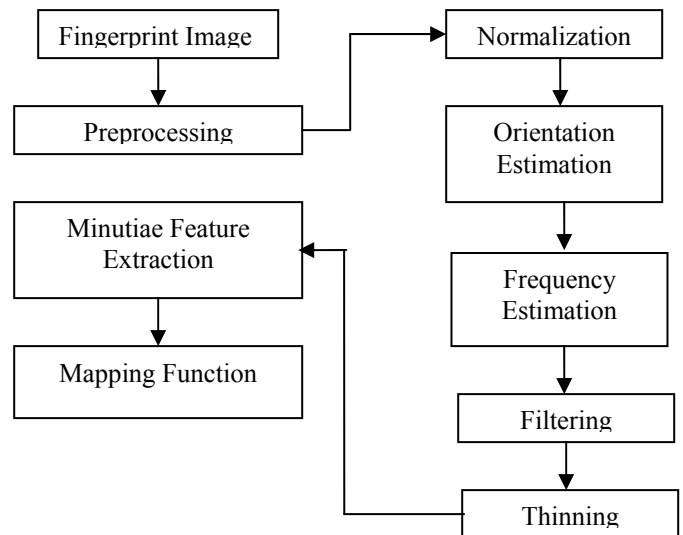
Figure 2 Steps involved in Extracting Feature Point

A. User Registration

This step is popularly known as Enrolment phase. In all the security system to enroll as a legitimate user in a service, a user must previously register with the service provider by ascertaining his/her identity with the provider. Therefore a scanner is used to scan the fingerprint of the user to reveal his/her identity for the first time. The finger print image thus obtained undergoes a series of enhancement steps. This is described in the following section of this proposed paper.

B. Fingerprint Enhancement

This is very important step in designing a security system for network security using biometrics. This step comprise of the subsequent processing on the obtained fingerprint image. As we all know a fingerprint is made of a series of ridges and furrows on the surface of the finger. This determines the uniqueness of the individuals fingerprint. No two fingerprints can have the same pattern of ridges and furrows. Minutiae points are local ridge characteristics that happen at either a ridge bifurcation or a ridge ending. The ridges hold the information of characteristic features obligatory for minutiae extraction therefore the quality of the ridge structures in a fingerprint image turns out to be an important characteristic. The obtained image is then subjected to image enhancement techniques to reduce the noise [11]. The following are the widely used image improvement techniques, normalization, orientation estimation, local frequency estimation, Gabor filtering, and thinning.

1 Normalization

The process of standardizing the intensity values in an image by adjusting the range of gray-level values so that it lies within a desired range of values is termed as "normalization". Moreover the ridge structures in the fingerprint are not affected as a result of this process. It is carried out to standardize the dynamic levels of variation in gray-level values that facilitates the processing of subsequent image

enhancement stages. Figure 3 shows a image of the fingerprint before and after normalization.

(a)   (b)

Figure 3. (a) Original Image (b) Image after normalization

2   Orientation Estimation

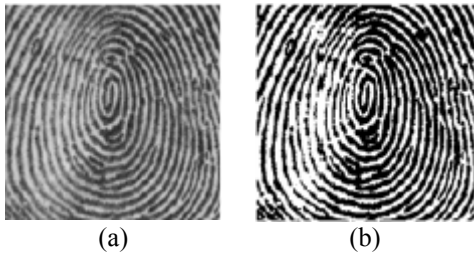The orientation estimation is an essential step in the enhancement process as the successive Gabor filtering stage relies on the local orientation in order to successfully enhance the fingerprint image. Figure 4 (a) and (b) illustrates the results of orientation estimation and smoothed orientation estimation of the fingerprint image respectively. In addition to the orientation image, another important parameter that is used in the construction of the Gabor filter is the local ridge frequency.
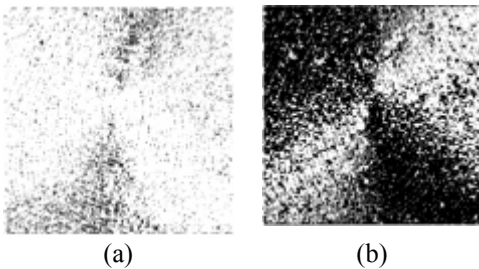

(a)   (b)

Figure 4. (a) Orientation Image (b) Smoothened Orientation Image

3   Gabor Filtering

Once the ridge orientation and ridge frequency information has been single-minded, these parameters are used to construct the even-symmetric Gabor filter. Gabor filters are employed because they have frequency-selective and orientation selective properties. These properties allow the filter to be tuned to give maximal response to ridges at a specific orientation and frequency in the fingerprint image. Therefore, a properly tuned Gabor filter can be used to effectively preserve the ridge structures while reducing noise. Figure 5 illustrates the results of using Gabor filter to a fingerprint image.

Figure   5   Filtered Image

4   Thinning

The concluding image enhancement pace typically performed former to minutiae extraction is thinning. Thinning is a morphological operation that successively erodes away the foreground pixels until they are one pixel wide. The application of the thinning algorithm to a fingerprint image preserves the connectivity of the ridge structures while forming a skeleton version of the binary image. This skeleton image is then used in the subsequent extraction of minutiae. Figure 6 shows the results of thinning to a fingerprint image.

Figure 6 Thinned Image

C.   Minutiae Feature Extraction

The next step is to extract the minutiae from the enhanced image. The most generally engaged technique of minutiae extraction is the Crossing Number (CN) concept [12]. This method engrosses the use of the skeleton image where the ridge flow pattern is eight-connected. The minutiae are extracted by scanning the local neighborhood of each ridge pixel in the image using a 3x3 window. The CN value is then computed, which is defined as half the sum of the differences between pairs of adjacent pixels in the eight-neighborhood. Figure 7 represents the list of minutiae in a fingerprint image.

Figure   7.   Minutiae extraction   on   a fingerprint image.

D.   Mapping Function

The coordinate system used to articulate the minutiae point locations of a fingerprint is a Cartesian coordinate system. The X and Y coordinate of the minutiae points are in pixel units. Angles are expressed in standard mathematical format, with zero degrees to the right and angles increasing in the counter-clockwise direction. Every minutia can be stored as a binary string. Each minutiae point can be recorded in 27 bits: 1 bit for the minutiae type, 9 bits each for minutiae X coordinate and Y coordinate, and 8 bits for the minutia angle. Thus, the binary representation of minutiae point is obtained. Suppose $M_i = (t_i, x_i, y_i, \theta_i)$ (i = 1 . . . n) are the all extracted minutiae for a fingerprint image. Then these minutiae points can be arranged

in a list from left to right by ascending X-co-ordinate, if equal by ascending Y-co-ordinate (first X, then Y) as follows:

$$Mi1Mi2 \cdots Min$$

The result of the feature extraction stage is what is called a minutia template (FP). An approximate range on the number of minutiae found at this stage is from 10 to 80. These are the different steps involved in designing a fingerprint based biometric authentication system for network security.

E. The Finger Print Hardening Protocol

There are two necessities for registration using Finger Print.
1. The user should obtain the biometric feature from his finger print using suitable image processing techniques as one mentioned in the previous section.
2. The minutia template should be encrypted with AES 128 bit symmetric cipher and is then transmitted to the server for storage in the database, so that it should not be possible for an outside attacker to determine the biometric feature by an exhaustive search either at the server side or by meet in the middle attack.

F. The Finger authentication Protocol

To initiate a request for service, user computes his FP1 = E_{AES}(FP), then the user sends the user ID along with FP1 to the server. In Lee et al.[16]'s protocol, the authority selects two large prime numbers p and q, where q|p-1. Let g be an element of order q in GF(p). Assume H(...) is a collision-free hash function with an output of q bits. The secret key of the sender S is $X_S \in Zq*$ and $Y_S = g^{Xs}$ mod p is the corresponding public key. Similarly, (X_R, Y_R) is the key pair of the receiver R, where $X_R \in Zq*$ and $Y_R \in gX_R$ mod p. The symbol "||" is the concatenate operator of strings. In this work, Li Gang's[17] protocol is adopted to implement the authentication protocol.
Let t1 and t2 be the minutiae template of FP1 and FP2,
Step 1. S chooses t, t1 $\in_R Zq*$ and computes r=g^t mod p, r_1=g^{t1} mod p and σ_1=H(r||T)X_S+$t_1 r_1$ mod q, where T is a time stamp, and then he sends (r, T, r1,σ1) to R;
Step 2. R checks whether $g^{\sigma 1} \equiv Y_s^{H(r||T)} r_1$ mod p. If not, R stops. Otherwise, R chooses $t_2 \in_R Z_q*$ and computes $r_2 = g^{t2}$ mod p and σ_2 H(r||T) X_R+$t_2 r_2$ mod q, and then he sends(r,T,r2, σ2) to S;
Step 3. S verifies whether $g^{\sigma 2} \equiv Y_R^{H(r||T)} r_2$ mod p.
If not, S stops. Otherwise, S computes
$$\sigma = H(M||T)X_S + tr \text{ mod } q,$$
$$k = (Y_R)^\sigma \text{ mod } p$$
and MAC=H(k||M||T||r_1||σ_1||r_2||σ_2). Finally, S sends MAC with M to R;
Step 4. R computes k'=$(Y_s^{H(M||T)} r^r$ mod p and verifies whether H(k'||M||T||r_1||σ_1|| r_2||σ_2)=MAC.
If the above equation holds, R accepts it. Otherwise, R rejects it and authentication becomes fail.

IV. PERFORMANCE MEASURES

This section of the paper explains the performance measures of our approach. The fingerprint processing has been done in

MATLAB 7. Some of the minutiae extracted from a sample finger print are shown in table 1. In the context of modern biometrics, these features, called fingerprint minutiae, can be captured, analyzed, and compared electronically, with correlations drawn between a live sample and a reference sample, as with other biometric technologies. There are two requirements for registration using Finger Print. The user should obtain the biometric feature from his finger print using appropriate image processing techniques as one mentioned in the previous section. The second is that the minutia template should be encrypted with AES 128 bit symmetric cipher and is then transmitted to the server for storage in the database, so that it should not be possible for an outside attacker to determine the biometric feature by an exhaustive search either at the server side or by meet in the middle attack.

Type	X	Y	Direction
1	35	117	2.93
1	50	83	2.95
0	19	57	2.80
0	23	135	0.27

Table 1.List of Minutiae

where 1 represent ridge ending point and 0 represent isolated point in a fingerprint image. Thus, the minutia can be expressed as a 4-vector with its elements in order, the type t, the X and Y coordinates (x, y), and the direction θ (Angle value is a non-negative value between 0 and 179, in units of degree) as shown in table 1. If each minutia is stored with type (1 bit), location (9 bits each for x and y), and direction (8 bits), then each will require 27 bits and the template will require up to 270 bytes. Then this binary representation is mapped on to a finger print hardening protocol for the generation of strong secret. The performance measures obtained revealed that the proposed method effectively provides network security. Therefore it can be directly applied to fortify existing standard single-server biometric based security applications. The analysis for the security of the protocol is based on the following assumptions (i) For a cyclic group G, generated by g, we are given g and g^n, $n \in N$, the challenge is to compute n. (ii) Given g, g^a, g^b, it is hard to compute g^{ab}. Clearly if these assumptions are not satisfied then C, an adversary, can gain access to the key gab. *A compromised session secret does not affect the security of the proposed deniable authentication protocol.*

The session secret can be derived from $k' \equiv Y_R^{XsH (M||T)+tr}$ mod *p*, where a random *t* is chosen independently from each session. If an attacker wants to forge the deniable information with the forged message *M'* by using the compromised session *k*, the receiver will derive a different session secret from the forged information. This is because that the message and its corresponding session secret are interdependent. Thereby, a compromised session secret does not affect the security of other sessions.

V. CONCLUSION

This paper proposes an approach for network security using biometrics. Biometric systems are commonly used to control

access to physical assets (laboratories, buildings, cash from ATMs, etc.) or logical information (personal computer accounts, secure electronic documents, etc). The human biometrics like hand geometry, face, fingerprint, retina, iris, DNA, signature and voice can be effectively used to ensure the network security. In biometric cryptosystems, a cryptographic key is generated from the biometric template of a user stored in the database in such a way that the key cannot be revealed without a successful biometric authentication. In this system, the ideas in the areas of image processing technique are reused to extract the minutiae from biometric image. The preprocessing techniques mentioned in this paper play an important role in improving the performance of the proposed biometric based network security system. The performance measures obtained revealed that the proposed method effectively provides network security. Therefore it can be directly applied to fortify existing standard single-server biometric based security applications. The future works rely on improving the network security by making use of cancelable biometrics and multimodal biometrics in the proposed authentication system.

REFERENCES

[1] Rajeswari Mukesh, A. Damodaram, and V. Subbiah Bharathi, "Finger Print Based Authentication and Key Exchange System Secure Against Dictionary Attack," IJCSNS International Journal of Computer Science and Network Security, Vol. 8, no. 10, pp. 14-20, 2008.

[2] T. Gunasekaran, and C. Parthasarathy, "Biometrics in Network Security," International Journal of Computer Network and Security (IJCNS), vol. 1, no. 1, pp. 36-42, 2006.

[3] Mahfuzur Rahman, and Prabir Bhattacharya, "Secure Network Communication Using Biometrics," IEEE International Conference on Multimedia and Expo (ICME'01), p. 52, 2001.

[4] Yunsu Chung, Kiyoung Moon, and Hyung-Woo Lee, "Biometric Certificate Based Biometric Digital Key Generation with Protection Mechanism," Frontiers in the Convergence of Bioscience and Information Technologies, pp. 709-714, 2007.

[5] Sandip Dutta, Avijit Kar, N. C. Mahanti, and B. N. Chatterji, "Network Security Using Biometric and Cryptography," Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 38-44, 2008.

[6] O. S. Benavente, and R. Piccio-Marchetti, "Authentication services and biometrics: network security issues," 39th Annual 2005 International Carnahan Conference on Security Technology, 2005. CCST '05, pp. 333-3336, 2005.

[7] Suriza Ahmad Zabidi, and Momoh-Jimoh E. Salami, "Design and Development of Intelligent Fingerprint-Based Security System," Knowledge-Based Intelligent Information and Engineering Systems, Book Chapter on Springer link, vol. 3214, pp. 312-318, 2004.

[8] Ronald G. Wolak, "Network Security: Biometrics - The Password Alternative," School of Computer and Information Sciences, 1998.

[9] Umut Uludag, Sharath Pankanti, Salil Prabhakar, and Anil K. Jain "Biometric Cryptosystems Issues and Challenges" Proceedings of the IEEE 2004.

[10] P. Arul, and Dr. A. Shanmugam, "Generate A Key for AES Using Biometric for VOIP Network Security," Journal of Theoretical and Applied Information Technology, pp. 107-112, 2009.

[11] L. Hong, Y. Wan, and A. Jain, "Fingerprint Image Enhancement: Algorithm and Performance Evaluation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no.8, pp.777-789, 1998.

[12] S. Kasaei, and B. Boashash, "Fingerprint feature extraction using block-direction on reconstructed images," In IEEE region TEN Conference on digital signal Processing applications, TENCON, pp. 303– 306, 1997.

[13] N. K. Ratha, J. H. Connell, and R. M. Bolle "Enhancing security and privacy in biometrics based authentication systems", IBM Systems Journal, vol. 40, pp. 614-634, 2001.

[14] Alexander P. Pons , and Peter Polak, "Understanding user perspectives on biometric technology," Communications of the ACM, vol. 51, no. 9, pp. 115-118, September 2008.

[15] "Biometrics Security Considerations," Systems and Network Analysis Center Information Assurance Directorate, www.nsa.gov/snac.

[16] W. B. Lee, C. C. Wu, and W. J. Tsaur, "A Novel Authentication Protocol Using Generalized ElGamal Signature Scheme", Information Sciences, 177, 2007, pp.1376-1381.

[17] Li Gang1, Xin Xiangjun, Li Wei, "An Enhanced Deniable Authentication Protocol," International Conference on Computational Intelligence and Security, Jan 2008

K. Saraswathi received her B.Sc., and M.C.A., from Avinashilingam University, Coimbatore, TamilNadu, in 1993 and 1996 respectively. She obtained her M.Phil degree from Bharathiar University, Coimbatore, TamilNadu, in the year 2003. Currently she is working as Assistant Professor, Department of Computer Science, Government Arts College, Udumalpet. She has the long experience of teaching Post graduate and Graduate Students. She is currently pursuing her Research in the area of Crypto Systems under Mother Teresa University, Kodaikanal, TamilNadu. Her area of interest includes Biometrics, Cryptography, Network Security, Machine Learning and Artificial Intelligence. She has Co-authored a text book on 'C' published by Keerthi Publications. She has presented her publications in various national conferences. She is a member of various professional bodies.

Dr. R. Balasubramanian was born in 1947 in India. He obtained his B.Sc., and M.Sc., degree in Mathematics from Government Arts College, Coimbatore, TamilNadu, in 1967 and PSG Arts College, Coimbatore, TamilNadu, in 1969 respectively. He received his Ph.D., from PSG College of Technology, Coimbatore, TamilNadu, in the year 1990. He has published more than 15 research papers in national and international journals. He has been serving engineering educational service for the past four decades. He was formerly in PSG College of Technology, Coimbatore as Assistant Professor in the Department of Mathematics and Computer Applications. He served as Associate Dean of the Department of Computer Applications of Sri Krishna College of Engineering and Technology, Coimbatore. Currently taken charge as Dean Academic Affairs at PPG Institute of Technology, Coimbatore, before which he was a Dean Basic Sciences at Velammal Engineering College, Chennai. He has supervised one PhD thesis in Mathematics and supervising four doctoral works in Computer Applications. His mission is to impart quality, concept oriented education and mould younger generation.

He is member of the board of studies of many autonomous institutions and universities. He was the principal investigator of UGC sponsored research project. He is a referee of an international journal on mathematical modeling. He has authored a series of books on Engineering Mathematics and Computer Science. He is a life member of many professional bodies like ISTE, ISTAM and CSI.

A New Region based Group Key Management Protocol for MANETs

N. Vimala
Senior Lecturer, Department of Computer Science
CMS College of Science and Commerce
Coimbatore, India**.**
vimalarmd@rediffmail.com

Dr. R. Balasubramanian
Dean Academic Affairs
PPG Institute of Technology
Coimbatore, India.
ramamurthybala2@gmail.com

Abstract-Key management in the ad hoc network is a challenging issue concerning the security of the group communication. Group key management protocols can be approximately classified into three categories; centralized, decentralized, and distributed. The most suitable solution to provide the services like authentication, data integrity and data confidentiality is the establishment of a key management protocol. This paper proposes an approach for the design and analysis of region-based key management protocols for scalable and reconfigurable group key management in Mobile Ad Hoc Networks (MANETs). Most of the centralized key management protocols arises an issue on data security on group communication. The proposed region-based group key management protocol divides a group into region-based subgroups based on decentralized key management principles. This region-based group key management protocols deal with outsider attacks in MANETs to preserve the security properties. A performance model to evaluate the network traffic cost generated for group key management in the proposed region-based protocol for MANETs is developed. Cost for joining or leaving the group and the cost for group communication are considered in evaluating the performance of the proposed region-based group key management scheme.

Keywords- Cluster Head, Group Key, Key Management Protocol, Mobile Ad Hoc Networks (MANETs), Region-based, and Rekeying.

I. INTRODUCTION

Generally, an ad hoc network is an assortment of independent nodes that communicate with each other, most regularly using a multi-hop wireless network. Nodes do not inevitably know each other and come together to form an ad hoc group for some particular reason. Key distribution systems typically involve a trusted third party (TTP) that acts as an intermediary between nodes of the network. A node in an ad hoc network has straight connection with a set of nodes, called neighboring nodes, which are in its communication range. The number of nodes in the network is not essentially preset. New nodes may join the network while existing ones may be compromised or become un-functional [1]. Key management in the ad hoc network is a challenging issue concerning the security of the group communication. Group key management protocols can be approximately classified into three categories; centralized, decentralized, and distributed [2].

MANET is one where there is no predetermined infrastructure such as base stations or mobile switching centers. Mobile nodes that are within each other's radio range communicate directly by means of a wireless network, whereas those far apart rely on other nodes to act as routers to relay its messages [3]. The most suitable solution to provide

the services among which authentication, data integrity and data confidentiality is the establishment of a key management protocol. This protocol is liable for the generation and the distribution of the traffic encryption key (TEK) to all the members of a group. This key is used by the source to encrypt multicast data and by the receivers to decrypt it. Therefore only legitimate members are able to receive the multicast flow sent by the group source [4]. The elemental security services provided by every key management system are key synchronism, secrecy, freshness, independence, authentication, confirmation, forward and backward secrecy [7].

Clustering is the concept of dividing the multicast group into a number of sub-groups. Each sub-group is managed by a local controller (LC), accountable for local key management within its cluster. Furthermore, not many solutions for multicast group clustering did think about the energy problem to realize an efficient key distribution process, whereas energy constitutes a foremost concern in ad hoc environments [5] [6]. The group key is generated by the cluster head and communicated to other members through a secure channel that uses public key cryptography [14]. Clusters may be used for achieving different targets [8]. Some of them are clustering for transmission management, clustering for backbone formation and clustering for routing efficiency. Group key management must be opposing to an extensive range of attacks by both outsiders and rouge members. In addition, group key management must be scalable, i.e., their protocols should be efficient in resource usage and able to decrease the effects of a membership change.

This paper proposes an approach for the design and analysis of region-based key management protocols for scalable and reconfigurable group key management in MANETs. This region-based group key management protocols deal with outsider attacks in MANETs to preserve the security properties. A performance model to evaluate the network traffic cost generated for group key management in the proposed region-based protocol for MANETs is developed.

The remainder of this paper is structured as follows. Section 2 of this paper discusses some of the earlier proposed cluster based group key management techniques. Section 3 describes our proposed method of new region based group key management protocol for MANETs. Section 4 explains the

performance evaluation of the proposed approach and section 5 concludes the paper with fewer discussions.

II. BACKGROUND STUDY

Key management is an indispensable part of any secure communication. Most cryptosystems rely on some underlying secure, robust, and efficient key management system. This section of the paper discusses some of the earlier proposed key management schemes for secure group communication in wireless ad hoc networks.

Maghmoumi et al. in [9] proposed a cluster based scalable key management protocol for Ad hoc networks. Their proposed protocol is based on a new clustering technique. The network is partitioned into communities or clusters based on affinity relationships between nodes. In order to ensure trusted communications between nodes they proposed two types of keys generated by each cluster head. The protocol is adaptive according to the limitation of the mobile nodes battery power and to the dynamic network topology changes. Their proposed approach of clustering based scalable key management protocol provided secured communications between the nodes of the Ad hoc networks.

A key management scheme for secure group communication in MANETs was described by Wang et al. in [10]. They described a hierarchical key management scheme (HKMS) for secure group communications in MANETs. For the sake of security, they encrypted a packet twice. They also discussed group maintenance in their paper in order to deal with changes in the topology of a MANET. Finally, they carried out a performance analysis to compare their proposed scheme with other conventional methods that are used for key management in MANETs. The results showed that their proposed method performed well in providing secure group communication in MANETs.

George et al. in [11] projected a framework for key management that provides redundancy and robustness for Security Association (SA) establishment between pairs of nodes in MANETs. They used a modified hierarchical trust Public Key Infrastructure (PKI) model in which nodes can dynamically assume management roles. Moreover they employed non-repudiation through a series of transactions checks to securely communicate new nodes information among Certificate Authorities (CAs). They assumed that nodes could leave and join the network at any time. Nodes could generate their own cryptographic keys and were capable of securing communication with other nodes. In order to balance the flexibility and increased availability of the Key Management Scheme (KMS), security was provided by introducing two concepts in addition to revocation and security alerts: non-repudiation and behavior grading. The KMS sustained sufficient levels of security by combining node authentication with an additional element, node behavior. A behavior grading scheme required each node to grade the behavior of other nodes.

A new group key management protocol for wireless ad hoc networks was put forth by Rony et al. in [12]. They put forth an efficient group key distribution (most commonly known as group key agreement) protocol which is based on multi-party Diffie-Hellman group key exchange and which is also password-authenticated. The fundamental idea of the protocol is to securely construct and distribute a secret session key, 'K,' among a group of nodes/users who want to communicate among themselves in a secure manner. The proposed protocol starts by constructing a spanning tree on-the-fly involving all the valid nodes in the scenario. It is understood, like all other protocols that each node is distinctively addressed and knows all its neighbors. The password 'P' is also shared among each valid member present in the scenario. This 'P' helps in the authentication process and prevents man-in-the-middle attack. Unlike many other protocols, the proposed approach does not need broadcast/multicast capability.

Bechler et al. in [13] described cluster-based security architecture for Ad hoc networks. They proposed and estimated a security concept based on a distributed certification facility. A network is separated into clusters with one special head node for each cluster. These cluster head nodes carry out administrative functions and shares a network key among other members of the cluster. Moreover the same key is used for certification. In each cluster, exactly one distinguished node–the cluster head (CH)–is responsible for establishing and organizing the cluster. Clustering is also used in some routing protocols for ad hoc networks. Decentralization is achieved using threshold cryptography and a network secret that is distributed over a number of nodes. The architecture addresses problems of authorization and access control, and a multi-level security model helps to adjust the complexity to the capabilities of mobile end systems. Based upon their authentication infrastructure, they provided a multi level security model ensuring authentication, integrity, and confidentiality.

A scalable key management and clustering scheme was proposed by Jason et al. in [15]. They projected a scalable key management and clustering scheme for secure group communications in ad hoc networks. The scalability problem is solved by partitioning the communicating devices into subgroups, with a leader in each subgroup, and further organizing the subgroups into hierarchies. Each level of the hierarchy is called a tier or layer. Key generation, distribution, and actual data transmissions follow the hierarchy. The Distributed Efficient Clustering Approach (DECA) provides robust clustering to form subgroups, and analytical and simulation results demonstrate that DECA is energy-efficient and resilient against node mobility. Comparing with most other schemes, their approach is extremely scalable and efficient, provides more security guarantees, and is selective, adaptive and robust.

Apart from the above mentioned numerous researches have been conducted in the field of cluster-based group key management for mobile ad hoc networks (MANETs).

III. A New Region Based Group Key Management for MANETs

The proposed region-based group key management protocol divides a group into region-based subgroups based on decentralized key management principles using Weighted Clustering Algorithm (WCA). This partitioning of region into subgroups improves scalability and efficiency of the key management scheme in providing a secure group communication. Figure 1 shows the partitioning of region into subgroups on the basis of decentralized key management principles [16, 18]. It is assumed that each member of the group is equipped with Global Positioning System (GPS) and therefore each one knows its location as it moves across the regions. For secure group communications, all members of a group share a secret group key, K_G. In addition to ensure security in communication between the members of each subgroup all the members of the subgroups in the region 'i' hold a secret key K_{Ri}. This shared secret key is generated and managed by a distributed group key management protocol that enhances robustness. This region-based group key management protocol will function at the optimal regional size recognized to reduce the cost of key management in terms of network traffic.
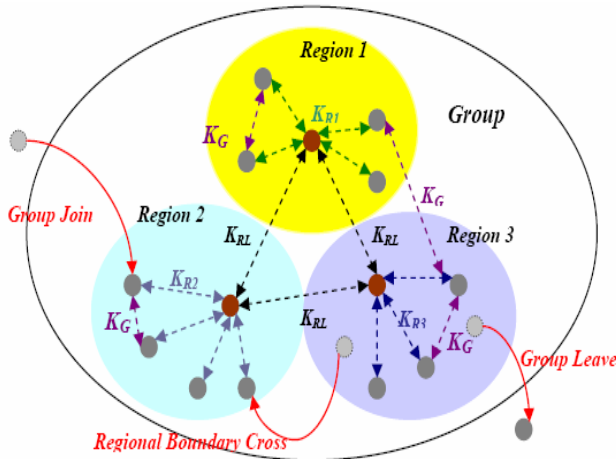

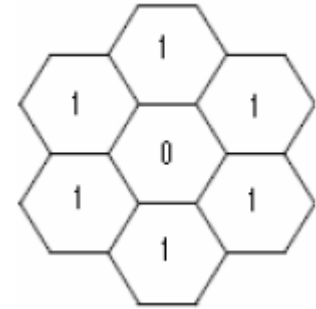Figure 1 Region-based Group Key Management

The average number of nodes in the system is $N=\lambda_p A$, where λ_p denotes the node density of the randomly distributed nodes and A indicates the operational area with radius 'r'. The random distribution of nodes is according to a homogeneous spatial Poisson process. The nodes can join or leave a group at any point of time. A node may leave a group at any time with rate μ and may rejoin any group with rate λ. Therefore, the probability that a node is in any group is $\lambda/(\lambda+\mu)$ and the probability that it is not in any group is $\mu / (\lambda+\mu)$. Let A_J and A_L be the aggregate join and leave rates of all nodes, respectively. Then, A_J and A_L, can be calculated as follows,

$$A_J = \lambda \times N \times \frac{\mu}{(\lambda + \mu)}$$

$$A_L = \mu \times N \times \frac{\lambda}{\lambda + \mu}$$

Nodes in a group must satisfy the forward/backward secrecy, confidentiality, integrity and authentication requirements for secure group communications in the presence of malicious outside attackers. The important requirement for secure group communication is reliable transmission. This can be achieved by using acknowledgement (ACK) packets and packet retransmission upon timeout. Hexagon is used to model a region [17]. Let R(n) denote the number of regions (i.e. $3n^2 + 3n + 1$) in the operational area. For n=3, the number of regions in the operational area is 37, for n=2 and n=1, the number of regions in the operational area are 19 and 7 respectively. Figure 2 shows the representation of the regions in the operational area for n=1, 2, and 3.

n=1, Number of Regions=7
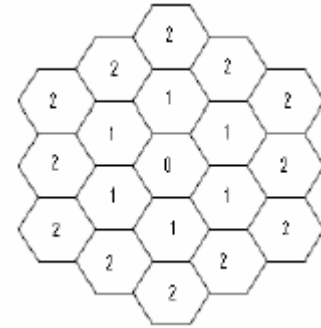
n=2, Number of Regions=19
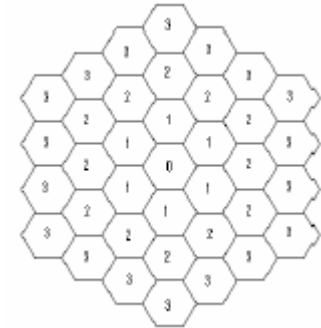
n=3, Number of Regions=37

Figure 2. Representation of Regions in operational area

A. Protocol Description

This describes the working of our proposed region-based group key management for MANETs.

1. Bootstrapping

In this initial bootstrapping process, a node within a region can take the responsibility of a regional "leader" to carry out Group Diffie Hellman (GDH). If there are multiple initiators, then the node with the smallest id will prevail as the leader and will implement GDH to completion to generate a regional key. Once a leader is generated in each region, all leaders in the group will execute GDH to agree on a secret leader key,

K_{RL}, for secure communications among leaders. The group key K_G can be generated using the following, K_G = MAC (K_{RL}, c), where MAC is a cryptographically secure hash function, K_{RL} is the leader key used as the secret key to MAC, and c is a fresh counter which will be incremented whenever a group membership event occurs. The generated group key K_G is then disseminated among the group members by the group leader. This group key provides secure group communication across regions.

2. Key Management

The next important task is managing the generated key. These shared secret keys at the subgroup (regional), leader, group levels may be rekeyed to preserve secrecy in response to events that occur in the system. Therefore, whenever there occur a change in the leader of the group, the leader key, K_{RL} is rekeyed. The regional key (K_R) is rekeyed whenever there is a regional membership change, including a local member group join/leave, a node failure, a local regional boundary crossing, and a group merge or partition event.

3. View Management

In addition to maintaining secrecy, the proposed region-based key management protocol also allows membership consistency to be maintained through membership views. Three membership views can be maintained by various parties: (a) Regional View (RV) contains regional membership information including regional (or subgroup) members' ids and their location information, (b) Leader View (LV) contains leaders' ids and their location information, and (c) Group View (GV) contains group membership information that includes members' ids and their location information.

4. Weighted Clustering Algorithm (WCA)

Weighted Clustering Algorithm (WCA) [18] selects a cluster head according to the number of nodes it can handle, mobility, transmission power and battery power. To avoid communications overhead, this algorithm is not periodic and the cluster head election procedure is only invoked based on node mobility and when the current dominant set is incapable to cover all the nodes. To ensure that cluster heads will not be over-loaded a pre-defined threshold is used which indicates the number of nodes each cluster head can ideally support. WCA selects the cluster heads according to the weight value of each node. The weight associated to a node v is defined as:

$$W_v = w1\,\Delta v + w2\,D_v + w3\,M_v + w4\,P_v$$

The node with the minimum weight is selected as a cluster head. The weighting factors are chosen so that $w1 + w2 + w3 + w4 = 1$. M_v is the measure of mobility. It is taken by computing the running average speed of every node during a specified time T. Δv is the degree difference. Δv is obtained by first calculating the number of neighbors of each node. The result of this calculation is defined as the degree of a node v, d_v. To ensure load balancing the degree difference Δv is calculated as $|d_v - \delta|$ for every node v, where δ is a pre-defined threshold. The parameter D_v is defined as the sum of distances from a given node to all its neighbors. This factor is related to energy consumption since more power is needed for larger distance communications. The parameter P_v is the cumulative time of a node being a cluster head. P_v is a measure of how much battery power has been consumed. A cluster head consumes more battery than an ordinary node because it has extra responsibilities. The cluster head election algorithm finishes once all the nodes become either a cluster head or a member of a cluster head. The distance between members of a cluster head, must be less or equal to the transmission range between them. No two cluster heads can be immediate neighbors

B. Rekeying protocol

Additional to group member join/leave events which cause rekeying of the group key, mobility-induced events may also cause rekeying. Below described is the proposed region-based key management protocol for a MANET in response to events that may occur in the system.

1. Group Member Join

The node willing to join the group initiates the process by sending a message "hello" along with its id and location information. Neighboring nodes receiving the beacon forward the "hello" message to their regional leader. The regional leader authenticates the new nodes identity based on its public key. Then, the leader acts as a coordinator involving all subgroup members including the new node to execute GDH to generate a new regional key. The leader then updates the regional membership list, and broadcasts the regional membership list to members in the region. This results in rekeying of group key, K_G. The regional leader informs the newly joined member's information to all other leaders. All leaders then concurrently share out the new group key to members in their regions by encrypting the group key with their respective regional key K_R.

2. Group Member Leave

When a non-leader member, say B, leaves the group, it informs its leaving objective to its regional leader. When the leader receives the leaving intention message from B, it updates its regional view and propagates the updated regional view to its members. Since a group leave event originate a regional membership change event, a new regional key is generated by executing GDH and distributed to the regional members. Next, the leader informs the membership change information to all other leaders. After all leaders receive the information on the current leave event, they also broadcast the changed group view to all their members. Finally, all leaders separately regenerate a group key and dispense it to their analogous members by encrypting the group key with their respective regional key K_R.

3. Boundary crossing by a non leader member

If a non-leader member crosses a regional boundary, for example, from region i to region j, a regional membership change occurs in both regions i and j. Thus, the regional keys in the two involved regions are respectively rekeyed based on GDH and the members' regional views in these two regions are updated. Since the mobility event changes neither the leader view nor the group view, no leader or group view updates are necessary. No rekeying of the group key is needed because the member leaving a region (subgroup) is still a member of the group.

4. Boundary Crossing by a Leader Member

If a leader member crosses a regional boundary from i to j, there is a leadership change in addition to all operations considered in the event of boundary crossing by a non-leader member. Thus, as in the group member leave by a leader member event, a new leader in the departing region is elected, the leader key is rekeyed among all leaders, and the leader view is updated among all leaders.

5. Leader Election

A group leave, a boundary crossing, or a disconnection by a leader member prompts a fresh leader election in the involved region. A member in the region after missing its regional leader's beaconing message can commence the execution of GDH and WCA based on its regional using cluster head election procedure. If there are more than one leader invoking GDH, the member with the minimum weight wins and will carry out GDH to produce a new regional key K_R. The new leader then announces itself as a new leader in the region by broadcasting a beacon message "I-am-a-new-leader" along with the new regional view encrypted with the regional key K_R.

C. Group Communication Protocol

For typical group communication, we accept to use the publish/subscribe service. It is assumed that all members are interested in all published data by all members. Thus, all published data in each member are disseminated to all members whenever each node publishes its data. By taking two-level hierarchical key management structure, the published data in each node is broadcast to its members in the region, and then the leader receiving the published data distributes it to other leaders. After then, each leader broadcasts the published data to its members respectively. When all published data are disseminated to all members in this way, a group key is used to encrypt/decrypt the published data.

IV. PERFORMANCE ANALYSIS

The performance analysis helps identify the optimal regional size that will minimize the network traffic generated while satisfying security properties in terms of secrecy, availability and survivability. The cost metric used for measuring the proposed group key management protocol is the total network traffic per time unit incurred in response to group key management events including regional mobility induced, group join/leave, periodic beaconing, and group merge/partition events. To evaluate the performance of this proposed approach we discuss on group join/leave cost, and group communication cost.

A. Group Join/Leave Cost

This is the cost per time unit for handling group join or leave events. This cost also includes the cost caused by connection/disconnection events by group members.

$$C_{Join/Leave,i} = \left[A_J \times C_{Join,i}\right] + \left[A_L \times C_{Leave,i}\right]$$

Here A_J and A_L are the aggregate group join and leave rates of all members, respectively. A group join event requires the update of the regional view and the rekeying of the regional key in the region from which the join event is originated, the cost of which is C_{intra}, as well as the update of the group view and the rekeying of a group key, the cost of which is C_{group}, i.

$$C_{Join,i} = \left[C_{int\,ra}\right] + \left[C_{group,i}\right]$$

The cost for group leave event includes two cases, namely, when a non-leader member leaves and when a leader leaves the group. Thus, the cost for a group leave event is given as follows,

$$C_{leave,i} = C_{leave,i}^{non-leader} + C_{leave,i}^{leader}$$

B. Cost for Group Communication

It includes the cost of group communications between members. It is assumed that the publish/subscribe service is used to realize efficient group communications. For simplicity, it is assumed that all members are interested in all published data by all members, and the data are published in each node with the rate of λ_{pub}. Thus, the aggregate rate that data are published in each node is obtained as:

$$A_{pub} = N \times \left[\frac{\lambda}{\lambda + \mu}\right] \times \lambda_{pub}$$

Whenever each node publishes its data, the published data should be disseminated to all members. Taking advantage of our hierarchical key management structure, the published data can be distributed to all leaders first, and then each leader can broadcast them to its members in the region.

$$C_{GC,i} = A_{pub} \times \left((N_{region,i} \times M_{pub} \times H_{region}) + (M_{pub} \times H_{leader,i})\right)$$

Figure 3 (a) shows the comparison of number of regions and cost for group join/leave and cost for group communication. Similarly Figure 3 (b) represents the comparison of C_{GC} and $C_{join/leave}$ for no region and optimal region.

(a)

(b)
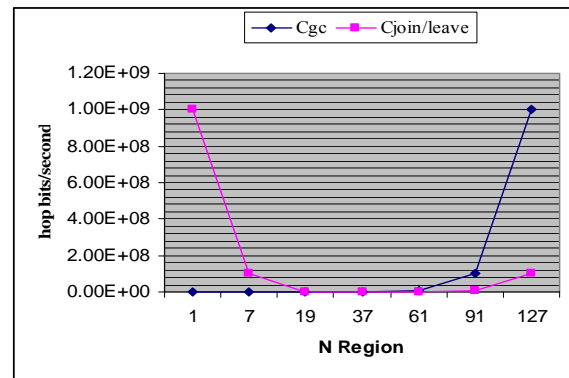
Figure 3 (a) shows the breakdown of C_{gc}, $C_{Join/Leave}$ Versus Number of regions and (b) C_{gc}, $C_{Join/Leave}$ under no region and optimal region

V. CONCLUSION

MANET is one where there is no predetermined infrastructure such as base stations or mobile switching centers. Key management in the ad hoc networ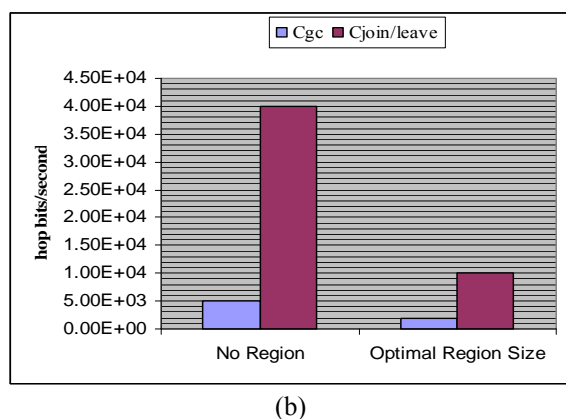k is a challenging issue concerning the security of the group communication. Group key management protocols can be approximately classified into three categories; centralized, decentralized, and distributed. This paper proposes an approach for the design and analysis of region-based key management protocols for scalable and reconfigurable group key management in MANETs. The proposed region-based group key management protocol divides a group into region-based subgroups based on decentralized key management principles. This region-based group key management protocols deal with outsider attacks in MANETs to preserve the security properties. In order to evaluate the network traffic cost generated for group key management of the proposed region-based protocol for MANETs a performance model is developed. Cost for joining or leaving the group and the cost for group communication are the parameters considered to investigate the performance of the proposed region-based group key management scheme.

REFERENCES

[1] A. Renuka, and K. C. Shet, "Cluster Based Group Key Management in Mobile Ad hoc Networks," IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 42-49, 2009.

[2] S. Rafaeli, and D. Hutchison, "A survey of key management for secure group communication," ACM Computing Surveys, vol. 35, no. 3, pp. 309–329, 2003.

[3] Hao Yang, Haiyun Luo, Fan Ye, Songwu Lu, and Lixia Zhang, "Security in mobile Ad-Hoc networks-Challenges and Solutions," IEEE Transactions on Wireless Communications, vol. 11, no. 1, pp. 38-47, 2004.

[4] Mohamed-Salah Bouassida, Isabelle Chrisment, and Olivier Festor, "Group Key Management in MANETs," International Journal of Network Security, vol. 6, no. 1, pp. 67-79, 2008.

[5] L. Lazos, and R. Poovendram, "Energy-aware secure multicast communication in Ad Hoc networks using geographical location information," in IEEE International Conference on Acoustics Speech and Signal Processing, pp. 201-204, 2003.

[6] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, "On the construction of energy-efficient broadcast and multicast trees in wireless networks," in INFOCOM 2000, pp. 585-594, 2000.

[7] Menezes, P. V. Oorschot, and S. A. Vanstone, "handbook of Applied Cryptography", CRC Press, New York, 1997.

[8] C. E. Perkins, "Ad hoc networking", Addison-Wesley Pub Co, 1st edition December 29, 2000.

[9] Chadi Maghmoumi, Hafid Abouaissa, Jaafar Gaber, and Pascal Lorenz, "A Clustering-Based Scalable Key Management Protocol for Ad Hoc Networks," Second International Conference on Communication Theory, Reliability, and Quality of Service, pp.42-45, 2009.

[10] Nen-Chung Wang, and Shian-Zhang Fang, "A hierarchical key management scheme for secure group communications in mobile ad hoc networks," Journal of Systems and Software, vol. 80, no. 10, pp. 1667-1677, 2007.

[11] George C. Hadjichristofi, William J. Adams, and Nathaniel J. Davis, "A Framework for Key Management in Mobile Ad Hoc Networks," International Journal of Information Technology, vol. 11, no. 2, pp. 31-61, 2006.

[12] Rony H. Rahman, and Lutfar Rahman, "A New Group Key Management Protocol for Wireless Ad-Hoc Networks," International Journal of Computer and Information Science and Engineering, vol. 2, no. 2, pp. 74-79, 2008.

[13] M. Bechler, H. -J. Hof, D. Kraft, F. Pählke, and L. Wolf, "A Cluster-Based Security Architecture for Ad Hoc Networks," Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM, vol. 4, pp. 2393-2403, 2004.

[14] Yi Jim Chen, Yi Ling Wang, Xian Ping Wu, and Phu Dung Le, "The Design of Cluster-based Group Key Management System in Wireless Networks," pp. 1-4, 2006.

[15] Jason H. Li, Renato Levy, Miao Yu, and Bobby Bhattacharjee, "A scalable key management and clustering scheme for ad hoc networks," Proceedings of the 1st international conference on Scalable information systems, 2006.

[16] Jin-Hee Cho, "Design and Analysis of QoS-Aware Key Management and Intrusion Detection Protocols for Secure Mobile Group Communications in Wireless Networks," Thesis submitted to the Faculty of the Virginia Polytechnic Institute and State University.

[17] J. W. Wilson, and I. R. Chen, "Performance Characteristics of Location-based Group Membership and Data Consistency Algorithms in Mobile Ad hoc Networks," International Journal of Wireless and Mobile Computing, vol. 1, no. 8, 2005.

[18] M. Chatterjee, S. K. Das, and D. Turgut, "An On-Demand Weighted Clustering Algorithm (WCA) for Ad hoc Networks," in proceedings of IEEE Globecom'00, pp. 1697–701, 2000.

N. Vimala received her B.Sc., (CS) from Avinashilingam Deemed University, Coimbatore, TamilNadu, in 1993. She obtained her M.Sc., (CS) and M.Phil degree from Bharathiar University, Coimbatore, TamilNadu, in the year 1995 and 2001 respectively. She is currently Senior Lecturer, Department of Computer Science, CMS College of Science and Commerce, Coimbatore, TamilNadu. She has the long experience of teaching Post graduate and Graduate Students. She has produced 43 M.Phil Scholars in various universities. Her area of interest includes Network Security, Database Management Systems, Object Oriented Programming and Artificial Intelligence. She is currently pursuing her Research in the area of Network Security under Mother Teresa University, Kodaikanal, TamilNadu. She is a member of various professional bodies.

Dr. R. Balasubramanian was born in 1947 in India. He obtained his B.Sc., and M.Sc., degree in Mathematics from Government Arts College, Coimbatore, TamilNadu, in 1967 and PSG Arts College, Coimbatore, TamilNadu, in 1969 respectively. He received his Ph.D., from PSG College of Technology, Coimbatore, TamilNadu, in the year 1990. He has published more than 15 research papers in national and international journals. He has been serving engineering educational service for the past four decades. He was formerly in PSG College of Technology, Coimbatore as Assistant Professor in the Department of Mathematics and Computer Applications. He served as Associate Dean of the Department of Computer Applications of Sri Krishna College of Engineering and Technology, Coimbatore. Currently taken charge as Dean Academic

Affairs at PPG Institute of Technology, Coimbatore, before which he was a Dean Basic Sciences at Velammal Engineering College, Chennai. He has supervised one PhD thesis in Mathematics and supervising four doctoral works in Computer Applications. His mission is to impart quality, concept oriented education and mould younger generation.

He is member of the board of studies of many autonomous institutions and universities. He was the principal investigator of UGC sponsored research project. He is a referee of an international journal on mathematical modeling. He has authored a series of books on Engineering Mathematics and Computer Science. He is a life member of many professional bodies like ISTE, ISTAM and CSI.

Automated Rapid Prototyping of TUG Specifications Using Prolog for Implementing Atomic Read/ Write Shared Memory in Mobile Ad Hoc Networks.

Fatma Omara[# 1], Said El Zoghdy[*2], Reham Anwer[*3]

[#] *Information Systems and Computers Faculty - Cairo University-Egypt.*

[1]Fatma_omara@hotmail.com

[*] *Science Faculty – Menufiya University- Egypt.*

[2] Elzoghdy@yahoo.com

[3] rehamteacher@yahoo.com

Abstract: **Rapid prototyping has been used for exploring vague user requirements in the front-end of the software life cycle. Automated rapid prototyping may reduce cost of prototyping and the time of developing it .One automated rapid prototyping technique is the direct execution of a specification. Direct execution of a specification has the benefits of quick construction of the prototype, direct support for formal specification, and quick response to the specification changes. However existing formal specification languages still have difficulties in specifying software systems such as non functional behavior of the systems. For non-executable formal specification languages, a prototype may be derived from the specification via software transformations. This approach to rapid prototyping uses a formal specification language to automatically generate a prototype in Prolog via a set of software transformation rules. Because there is a direct correspondence between the language and Prolog, the transformation is mechanical and straight forward. Specifiers can concentrate on generating the prototype without the distraction of transforming one notation into another. This formal specification language may not provide enough abstractions for provide enough abstraction for prototyping some particular features of systems. Therefore, this approach is designed to support the derived prototype to be extended or modified in a modular manner. The specification is written in modules in terms of the language patterns that support module independence, the prototype is then derived in a modular way that supports the ease of modifications to the prototype. The software transformation rules used for the derivation of prototypes in Prolog are presented. In this paper, we apply this specification on the implementation for atomic object Read/Write shared memory in mobile ad hoc network.**

Keywords: Rapid Prototyping, TUG language, Prolog, Mobile Ad Hoc Networks.

I. INTRODUCTION

Tree Unified with Grammar (TUG) was developed to support a system to be developed through an integration of conventional software development, operational specification, rapid prototyping via software transformations, software reuse, and analysis of specifications and programs via testing and proofs. The language integrates various software development paradigms into a coherent whole to fit specific needs of developing organizations. This language improves the reusability of formal specifications in the following ways [1]:

(1) A developer can run a TUG specification as a prototype to study its behavior due to its executability at the front-end of the software life cycle

(2) A developer can easily write a parametric program corresponding to its parametric.

The idea of prototyping via Software transformations isn't new. However, as automatic rapid prototyping approach should avoid a proto- type to be rederived from scratch whenever there is a change in the specification. Also, the automated approach should allow developers to easily extend the functions of the prototype manually in case the specification language doesn't support abstractions for features needed for demonstration. A rapid prototyping approach via Software transformations is presented to achieve this goal. User requirement are first written into specification in TUG. The specification needs not necessarily be complete, precise and correct corresponding to the user requirements at the first attempt. However, the specification should comply with the syntax of the language in order to be further processed. Next, a prototype in prolog is automatically derived from the specification. Via software transformations. The prototype is then exercised by the specifier and the use to clarify the user requirements in the front- end of the Software life cycle.

II. THE GEOQUORUM-APPROACH

In this paper the GeoQuorums approach has presented for implementing atomic read/write shared memory in mobile ad hoc networks. This approach is based on associating abstract atomic objects with certain geographic locations. It is assumed that the existence of Focal Points, geographic areas that are normally "populated" by mobile nodes. For example: a focal point may be a road Junction, a scenic observation point [2]. Mobile nodes that happen to populate a focal point participate in implementing a shared atomic object, using a replicated state machine approach. These objects, which are called focal point objects, are prone to occasional failures when the corresponding geographic areas are depopulated. The Geoquorums algorithm uses the fault-prone focal point objects to implement atomic read/write operations on a fault-tolerant virtual shared object. The Geoquorums algorithm uses a quorum- based strategy in which each quorum consists of a set of focal point objects. The quorums are used to maintain the consistency of the shared memory and to tolerate limited failures of the focal point objects, which may be caused by depopulation of the corresponding geographic areas. The mechanism for changing the set of quorums has presented, thus improving efficiency [2]. Overall, the new Geoquorums algorithm efficiently implements read/write operations in a highly dynamic, mobile network. In this chapter, a new approach to designing algorithms for mobile ad hoc networks is presented. An ad hoc network uses no pre-existing infrastructure, unlike cellular networks that depend on fixed, wired base stations. Instead, the network is formed by the mobile nodes themselves, which co-operate to route communication from sources to destinations. Ad hoc communication networks are by nature, highly dynamic. Mobile nodes are often small devices with limited energy that spontaneously join and leave the network. As a mobile node moves, the set of neighbors with which at can directly communicate may change completely. The nature of ad hoc networks makes it challenging to solve the standard problems encountered in mobile computing, such as location management using classical tools. The difficulties arise from the lack of a fixed infrastructure to serve as the backbone of the network. In this chapter developing a new approach that allows existing distributed algorithm to be adapted for highly dynamic ad hoc environments one such fundamental problem in distributed computing is implementing atomic read/ write shared memory [2]. Atomic memory is a basic service that facilitates the implementation of many higher level algorithms. For example: one might construct a location service by requiring each mobile node to periodically write its current location to the memory. Alternatively, a shared memory could be used to collect real – time statistics. The problem of implementing atomic read/write memory is divided into two parts; **first**, we define a static system model, the focal point object model that associates abstract objects with certain fixed geographic locales. The mobile nodes implement this model using a replicated state machine approach. In this way, the dynamic nature of the ad hoc network is masked by a static model. Moreover it should be noted that this approach can be applied to any dynamic network that has a geographic basis. **Second**, an algorithm is presented to implement read/write atomic memory using the focal point object model. The implementation of the focal point object model depends on a set of physical regions, known as focal points [2].The mobile nodes within a focal point cooperate to simulate a single virtual object, known as a focal point object. Each focal point supports a local broadcast service, LBcast which provides reliable, totally ordered broadcast. This service allows each node in the focal point to communicate reliably with every operation completely. The focal broadcast service is used to implement a type of a replicated state machine, one that tolerates joins and leaves of mobile nodes. If a focal point becomes depopulated, then the associated focal point object fails. (Note that it doesn't matter how a focal point becomes depopulated, be it as a result of mobile nodes failing, leaving the area, going to sleep. etc. Any depopulation results in the focal point failing). The Geoquorums algorithm implements an atomic read/write memory algorithm on top of the geographic abstraction, that is, on top of the focal point object model. Nodes implementing the atomic memory use a Geocast service to communicate with the focal point objects. In order to achieve fault tolerance and availability, the algorithm replicates the read/write shared memory at a number of focal point objects. In order to maintain consistency, accessing the shared memory requires updating certain sets of focal points known as quorums. An important aspect of this approach is that the members of our quorums are focal point objects, not mobile nodes. The algorithm uses two sets of quorums (I) **get-quorums** (II) **put- quorums** with property that every get-quorum intersects every put-quorum. There is no requirement that put-quorums intersect other put-quorums, or get-quorums intersect other get-quorums. The use of quorums allows the algorithm to tolerate the failure of a limited number of focal point objects. This algorithm uses a Global Position System (GPS) time service, allowing it to process write operations using a single phase, prior single-phase write algorithm made other strong assumptions, for example: relying either on synchrony or single writers. This algorithm guarantees that all read operations complete within two phases, but allows for some reads to be completed using a single phase: the atomic memory algorithm flags the completion of a previous read or write operation to avoid using additional phases, and propagates this information to various focal paint objects. As far as we know, this is an improvement on previous quorum based algorithms. For performance reasons, at different times it may be desirable to use different times it may be desirable to use different sets of get quorums and put-quorums [2]. For example: during intervals when there are many more read operations than write operations, it may be preferable to use smaller get- quorums that are well distributed, and larger put-quorums that are sparsely distributed. In this case a client can rapidly communicate with a get-quorum while communicating with a put – quorum may be slow. If the operational statistics change, it may be useful to reverse the situation.

A. Mathematical Notation for Geoquorums Approach

- I the totally- ordered set of node identifiers.
- $i_0 \in I$, a distinguished node identifier in I that is smaller than all order identifiers in I.
- S, the set of port identifiers, defined as $N^{>0} \times OP \times I$, Where OP= {get, put, confirm, recon- done}.
- O, the totally- ordered, finite set of focal point identifiers.
- T, the set of tags defined as $R^{\geq 0} \times I$.
- U, the set of operation identifiers, defined as $R^{\geq 0} \times S$.
- X, the set of memory locations for each $x \in X$:
  - V_x the set of values for x
  - $v_{0,x} \in V_x$, the initial value of X
- M, a totally-ordered set of configuration names
- $c_0 \in M$, a distinguished configuration in M that is smaller than all other names in M.
- C, totally- ordered set of configuration identifies, as defined as:

$$R^{\geq 0} \times I \times M$$

- L, set of locations in the plane, defined as $R \times R$

Fig.1 Notations Used in the Geoquorums Algorithm.

B. Variable Types for Atomic Read/Write object in Geoquorum Approach for Mobile Ad Hoc Network

The specification of a variable type for a read/write object in geoquorum approach for mobile ad hoc network is presented and a read/write object has the following variable type (see figure .2) .

Put/get variable type τ

State

Tag $\in$ T, initially< $0.i_0$>

Value $\in$ V, initially v_0

Config-id $\in$ C, initially< 0, i_0, c_0>

Confirmed-set C T, initially Ø

Recon-ip, a Boolean, initially false

Operations

Put (new-tag, new-value, new-config-id)

If (new-tag> tag) then

Value ⬚new-value

Tag ⬚ new-tag

If (new-config-id > config-id) then

Config-id ⬚ new-config-id

Recon-ip ⬚ true

C. Operation Manager

In this section the Operation Manger (OM) is presented, an algorithm built on the focal/point object Model. As the focal point Object Model contains two entities, focal point objects

and Mobile nodes, two specifications is presented , on for the objects and one for the application running on the mobile nodes [2] .

1) Operation Manager Client: This automaton receives read, write, and recon requests from clients and manages quorum accesses to implement these operations (see fig .3, fig.4, and fig.5). The Operation Manager (OM) is the collection of all the operation manager clients (OM$_i$, for all i in I).it is composed of the focal point objects, each of which is an atomic object with the put/get variable type.

Signature:
Input:
Write (Val)$_i$,val $\in$ V
read ( )$_i$
recon (cid)$_I$, cid $\in$ C
respond (resp) $_{obj, P,}$ resp$\in$ responses (τ), obj $\in$ O, P = <*,*, i> $\in$ S
geo-update (t,L)$_i$, t$\in$ $R^{\geq 0}$, L $\in$ L
output:
write-ack ( )$_i$
read-ack (val)$_i$,val $\in$ V
recon-ack (cid)$_i$, cid $\in$ C
invoke (inv) $_{obj, P,}$ inv $\in$ invocations (τ),obj $\in$ O, P = <*,*,i> $\in$ S
Internal:
read-2 ( )$_i$
recon-2 (cid)$_i$, cid $\in$ C
State:
Confirmed $\underline{C}$ T, a set of tag ids, initially Ø
Conf-id $\in$ C, a configuration id, initially <0, i_0, c_0>
recon- ip, a Boolean flag initially false
Clock $\in$ $R^{\geq 0}$, a time, initially 0
Ongoing ‾ invocations $\underline{C}$ O$\times$ S a set of objects and ports initially Ø
Current-port-number $\in$ N $^{>0}$, used to invoke objects, initially 1
Op, a record with the following components:
Type $\in$ {read, write, recon}, initially read
Phase $\in$ {{idle, get, put}, initially idle
Tag $\in$ T, initially <0, i_0>
Value $\in$ V, initially v_0
recon-ip , a Boolean flag, initially false
Recon-conf-id $\in$ C, a configuration id, initially <0, i_0, c_0>
Acc $\underline{C}$ 0, a set of data objects, initially Ø

Fig. 3 Operation Manager Client Signature and State for Node i in I, where τ is the Put/Get Variable Type.

Operation Manager Client Transitions
Output invoke (<get, config-id>) $_{obj,p}$
Preconditions:
P= <current-port-number, get, i>
<obj, p> $\notin$ ongoing-invocations
Obj $\notin$ op.acc

Op.phase =get
Config-id=conf-id
Effect:
Ongoing-invocations ⟵
Ongoing-invocations U {<obj,p>}
Output invoke (<put, tag, val, config-id>) $_{obj,p}$
Precondition:
P=<current-port-number, put, i>
<obj,p>∉ ongoing-invocations
Obj ∉ op.acc
Op.phase =put
tag=op.tag
Val=op.value
Config-id=conf-id
Effect:
Ongoing-invocations ⟵ Ongoing-invocations U
{<obj, p>}
Output invoke (<confirm, tag >) $_{obj, p}$
Precondition:
P=<k, confirm, i>
<obj,p>∉ ongoing-invocations
tag ∈ confirmed
Effect:
Ongoing-invocations ⟵
Ongoing-invocations U {<obj,p>}
Output invoke (<recon-done, config-id >) $_{obj,p}$
Precondition:
P=<k, recon-done,i>
<obj,p> ∈ ongoing-invocations
Recon-ip=false
Config-id=conf-id
Input respond (<get-ack, tag, val, confirmed, new-cid, new-rip>) $_{obj,p}$
Effect:
If(<current-port-number,get,i>=p)then
Op.acc ⟵ op.acc U {obj}
If (tag>op.tag) then
Op.tag ⟵ tag
Op.value ⟵ val
If (new-cid>conf-id) then
Conf-id ⟵ new-cid
Op.recon-ip ⟵ true
Recon-ip ⟵ new-rip
Else if (new-cid=conf-id) then
Recon-ip ⟵ recon-ip ∧ new-rip
If(confirm=true)then
Confirmed ⟵ confirmed U {tag}
Ongoing-invocations ⟵
Ongoing-invocations \ {<obj,p>}
Input respond (<put-ack, new-cid, new-rip>) $_{obj,p}$
Effect:
If (<current-port-number, put, i>=p) then
Op.acc ⟵ op.acc U {obj}
If (new-cid>conf-id) then
Conf-id ⟵ new-cid
Op.recon-ip ⟵ true
Recon-ip ⟵ new-rip
Else if (new-cid=conf-id) then
Recon-ip ⟵ recon-ip ∧ new-rip

Ongoing-invocations ⟵
Ongoing-invocations \ {<obj,p>}
Input respond (<confirm-ack>) $_{obj,p}$
Effect:
Ongoing-invocations ⟵
Ongoing-invocations \ {<obj,p>}
Input respond (<recon-done-ack>) $_{obj,p}$
Effect:
Ongoing-invocations ⟵
Ongoing-invocations \ {<obj,p>}

Fig.4 Operation Manager Client Invoke/Respond Transitions for node i

Operation Manager Client Transitions
Input write (val) $_i$
Effect:
Current-port-number⟵
Current-port-number +1
Op ⟵
<write, put,<clock,i>,Val, recon-ip,<0,i_0,c_0>, Ø>
Output write-ack ( ) $_i$
Precondition:
Conf-id=<time-stamp, pid, c>
If op.recon-ip then
√ C$^/$ ∈ M, ∋ P ∈ put-quorums(C$^/$): P ⊆ op.acc
Else
∃ P ∈ put-quorums(C): P ⊆ op.acc
Op.phase=put
Op.type=write
Effect:
Op.phase ⟵ idle
Confirmed ⟵ confirmed U {op.tag}
Input read ( ) $_i$
Effect:
Current-port-number ⟵
Current-port-number +1
Op ⟵ < read, get,$⊥,_⊤$, recon-ip, <0,i_0,c_0>, Ø>
Output read-ack (v) $_i$
Precondition:
Conf-id=<time-stamp, pid, c>
If op.recon-ip then
√ C$^/$ ∈ M, ∋ G ∈ get-quorums(C$^/$): G ⊆ op.acc
Else
∃ G ∈ get-quorums(C): G ⊆ op.acc
Op.phase=get
Op.type=read
Op.tag ∈ confirmed
v= op.value
Effect:
Op.phase ⟵ idle
Internal read-2( ) $_i$
Precondition:
Conf-id=<time-stamp, pid, c>
If op.recon-ip then
√ C$^/$ ∈ M, ∋ G ∈ get-quorums(C$^/$): G ⊆ op.acc
Else
∃ G ∈ get-quorums(C): G ⊆ op.acc
Op.phase=get
Op.type=read

Op.tag ∉ confirmed-set
 Effect:
Current-port-number ←
Current-port-number +1
Op.phase ← put
Op.recon.ip ← recon-ip
Op.acc ← Ø
Output read-ack (v)$_i$
Precondition:
Conf-id=<time-stamp, pid, c>
If op.recon-ip then
√ C′ ∈ M, ϶ P ∈ put-quorums(C′): P $\underline{C}$ op.acc
Else
϶ P ∈ put-quorums(C): P $\underline{C}$ op.acc
Op.phase=put
Op.type=read
v=op.value
Effect:
Op.phase ← idle
Confirmed ← confirmed U {op.tag}
Input recon (conf-name)$_i$
Effect:
Conf-id ← <clock,i, conf-name>
recon-ip ← true
Current-port-number ←
Current-port-number +1
Op ← < recon, get,⊥, ⊥, true,conf-id, Ø>
Internal recon-2(cid)$_i$
Precondition
√ C′ ∈ M, ϶ G ∈ get-quorums(C′): G $\underline{C}$ op.acc
√ C′ ∈ M, ϶ P ∈ put-quorums(C′): P $\underline{C}$ op.acc
Op.type=recon
Op.phase=get
cid=op.recon-conf-id
Effect
Current-port-number ←
Current-port-number +1
Op.phase ← put
Op.acc ← Ø
Output recon-ack(c)$_i$
Precondition
cid=op.recon-conf-id
cid= <time-stamp, pid, c>
϶ P ∈ put-quorums(C): P $\underline{C}$ op.acc
Op.type=recon
Op.phase=put
Effect:
If(conf-id=op.recon-conf-id)then
Recon-ip ← false
Op.phase ← idle
Input geo-update (t, L) $_i$
Effect:
Clock ← 1

Fig.5 Operation Manager Client Read/Write/Recon and Geo-update Transitions for Node

D. Focal Point Emulator Overview

The focal point emulator implements the focal point object Model in an ad hoc mobile network. The nodes in a focal point (i.e. in the specified physical region) collaborate to implement a focal point object. They take advantage of the powerful LBcast service to implement a replicated state machine that tolerates nodes continually joining and leaving .This replicated state machine consistently maintains the state of the atomic object, ensuring that the invocations are performed in a consistent order at every mobile node [3]. In this section an algorithm is presented to implement the focal point object model. the algorithm allows mobile nodes moving in and out of focal points, communicating with distributed clients through the geocast service, to implement an atomic object (with port set q=s)corresponding to a particular focal point. We refer to this algorithm as the Focal Point Emulator (FPE).fig .6 contains the signature and state of the FPE .the code for the FPE client is presented in fig.7. The FPE client has three basic purposes. First, it ensures that each invocation receives at most one response (eliminating duplicates).Second, it abstracts away the geocast communication, providing a simple invoke/respond interface to the mobile node[2] [3]. Third, it provides each mobile node with multiple ports to the focal point object; the number of ports depends on the atomic object being implemented. The remaining code for the FPE server is in fig .8.When a node enters the focal point, it broadcasts a join-request message using the LBcast service and waits for a response. The other nodes in the focal point respond to a join-request by sending the current state of the simulated object using the LBcast service. As an optimization, to avoid unnecessary message traffic and collisions, if a node observes that someone else has already responded to a join-request, and then it does not respond. Once a node has received the response to its join-request, then it starts participating in the simulation, by becoming active. When a node receives a Geocast message containing an operation invocation, it resends it with the lbcast service to the focal point, thus causing the invocation to become ordered with respect to the other LBcast messages (which are join-request messages, responses to join requests, and operation invocations).since it is possible that a Geocast is received by more than one node in the focal point ,there is some bookkeeping to make sure that only one copy of the same invocation is actually processed by the nodes. There exists an optimization that if a node observes that an invocation has already been sent with LBcast service, then it does not do so. Active nodes keep track of operation invocations in the order in which they receive them over the LBcast service. Duplicates are discarded using the unique operation ids. The operations are performed on the simulated state in order. After each one, a Geocast is sent back to the invoking node with the response. Operations complete when the invoking node with the response. Operations complete when the invoking node remains in the same region as when it sent the invocation, allowing the geocast to find it. When a node leaves the focal point, it re-initializes its variables [2] [3].A subtle point is to decide when a node should start collecting invocations to be applied to its replica of the object state. A node receives a snapshot of the state when it joins. However by the time the snapshot is received, it might be out of date, since there may have been some intervening messages from the LBcast service that have been received since the

snapshot was sent. Therefore the joining node must record all the operation invocations that are broadcast after its join request was broadcast but before it received the snapshot .this is accomplished by having the joining node enter a "listening" state once it receives its own join request message; all invocations received when a node is in either the listening or the active state are recorded, and actual processing of the invocations can start once the node has received the snapshot and has the active status. A precondition for performing most of these actions that the node is in the relevant focal point. This property is covered in most cases by the integrity requirements of the LBcast and Geocast services, which imply that these actions only happen when the node is in the appropriate focal point[2] [3].

Signature:

Input

Geocast-rcv (< invoke, inv, oid, Loc>) $_{obj,i}$,inv $\in$ invocations, oid $\in$ U,loc $\in$ L   (i.e. oid: object identifier , loc: location)

Lbcast-rcv (<Join-req, jid >) $_{obj,i}$, Jid $\in$ T (i.e. jid: join identifier)

Lbcast-rcv (<Join-ack, jid, v>) $_{obj,I}$, Jid $\in$ T, v$\in$ V

Lbcast-rcv (<invoke, inv, oid, loc>) $_{obj, i}$, inv $\in$ invocations, oid $\in$ U, loc $\in$ L

Geo-update (l,t)$_{obj,I}$,l $\in$ L ,t $\in$ R$^{>0}$

Output:

Geocast (<response, resp, oid, loc>) $_{obj,i}$, resp $\in$ Responses, oid$\in$ U , loc$\in$ L

Lbcast (< Join-req, Jid>) $_{obj, i}$, jid $\in$ T

Lbcast (< Join-ack, Jid, v>) $_{obj, i}$, jid $\in$ T, v$\in$ V

Lbcast (< invoke, inv, oid, loc>) $_{obj, i}$, inv $\in$ invocations, oid $\in$ U, loc $\in$ L

Internal:

Join ( ) $_{obj,i}$

Leave ( ) $_{obj,i}$

Simulate-op(inv)$_{obj,i}$, inv$\in$ invocations.

State:

Fp-location $\in$ 2^L, constant, locations defining the focal point under consideration

Clock $\in$ R$^{>0}$, the current time, initially

 0, updated by the geosensor.

 Location$\in$, L the current physical location, updater by the geosensor

Status $\in$ {idle, joining, listening, active}, initially active if node is in FP-location and idle otherwise.

Join- id $\in$ T, unique id for current join request, initially <0, i$_0$>.

Lbcast- queue, a queue of messages to be sent by the LBcast, initially Ø.

Geocast-queue, a queue of messages to be sent by the Geocast, initially Ø.

Answered-join-reqs set of ids of Join requests that have already been answered, initially Ø.

val $\in$ V, holds current value of the simulated atomic object, initially v$_0$.

Pending-ops, queue of operations waiting to be simulated, initially Ø.

Completed-ops, queue of operations that have been simulated, initially Ø.

Fig. 6 FPE server signature and state for node i and object obj of variable type τ = <V, v$_0$, invocations, responses, δ >

Signature:

Input:

Invoke (inv) $_{obj, p}$, inv $\in$ invocations, P $\in$ Q

Geocast –rcv (<response, resp, oid, loc>)$_{obj,i}$, resp $\in$ responses, oid $\in$ U

Geo-update (l,t)$_{obj,i}$, l $\in$ L , t E R $^{>0}$

Output:

Geocast (m) $_{obj,i}$, m $\in$ invoke×invocations×U ×L×L

respond (resp) $_{obj,p}$, resp $\in$ responses, p $\in$ Q

State:

Fp-location $\in$ 2^L, a constant, the locations of the focal point

Clock $\in$ R$^{\geq 0}$, the current time, initially 0, updated by geosensor

Location $\in$ L, the current physical location, updated by geosensor

ready-responses $\underline{C}$ Q×responses, a set of operation responses, initially Ø

Geocast-queue, a queue of messages to be geocast

ongoing-oids $\underline{C}$ U, a set of operation identifiers, initially Ø

Transitions:

Input invoke (inv)$_{obj, P}$

Effect:

New-oid⬜ <clock, p>

Enqueue (geocast-queue, < invoke, inv, new-oid, location, fp-location>)

Ongoing-oids⬜ ongoing-oids U {new-oid}

Input geocast-rcv (<response, resp, oid, loc>)$_{obj,i}$

Effect:

If (oid $\in$ ongoing –oids) then

<C, p> ⟵ oid

ready-responses⬜ ready-responses U {<p, resp>}

Ongoing-oids ⬜ Ongoing-oids \ {oid}

Input geo –update (l,t)$_{obj,i}$

Effect:

Location ⬜ L

Clock ← t

output geocast (m) $_{obj,i}$

Precondition:

Peek (geocast-queue) =m

Effect:

Dequeue (geocast-queue)

Output respond (resp) $_{obj,i}$

Pre conciliation:

< P, resp> ∈ ready-responses

Effect:

Ready-responses ← ready-responses \{<p, resp >}

Fig. 7 FPE client for client i and object obj of variable type τ = <V, v$_0$, invocations, responses, δ >

Focal Point Emulator Server Transitions

Internal join ( ) $_{obj,i}$

Precondition:

Location ∈ fp-location

Status=idle

Effect:

Join-id ←<clock, i>

Status← joining

Enqueue (lbcast-queue, <join-req, join-id>)

Input lbcast-rcv (< join-req, jid>) $_{obj,i}$

Effect:

If ((status=joining)) ∧ (jid=Join-id)) then

Status ⬚listening

If ((status=active))) ^ jid ∉ answered-join-reqs)) then

Enqueue (LBcast-queue,< join-ack, jid, val>)

Input Lbcast-rcv (<join-ack, jid, v>) $_{obj,i}$

Effect:

Answered-join-reqs ⬚ answered-join-reqs U {jid}

if ((status=listening) ∧ (jid =join-id)) then

Status ⬚ active

Val ⬚ v

Input geocast –rcv (< invoke, inv, oid, loc, fp-loc>) $_{obj,i}$

Effect:

If (fp-loc=fp-location) then

If (<inv, oid, loc>∉ pending-ops U completed ops) then

Enqueue (Lbcoast-queue, <invoke, inv, oid, loc>)

Input LBcast –rcv (< invoke, inv, oid, loc>) $_{obj,i}$

Effect:

If ((status=listening V active) ^

 (<inv, oid, loc>∉ pending-ops U completed-ops)) Then

Enqueue (pending-ops,<inv, oid, loc>)

Internal simulate-op (inv) $_{obj,i}$

Precondition:

Status=active

Peek (pending-ops) =<inv, oid, loc>

Effect:

(Val, resp)⬚ δ (inv, val)

 Enqueue (geocost- queue,< response, resp, oid, loc>)

Enqueue (completed-ops, Dequeue (pending-ops))

Internal leave ( ) $_{obj,i}$

Precondition:

Location ∉ fp-location

Status ≠ idle

Effect:

Status⬚ idle

Join-id⬚ <0,i$_0$>

Val ⬚ v$_0$

answered -join- reqs← Ø

Pending –ops ← Ø

Completed-ops ← Ø

Lbcast-queue ← Ø

Geocast-queue ← Ø

Output Lbcast (m) $_{obj,i}$

Precondition:

Peek (Lbcast-queue) =m

Effect:

Duqueue (Lbcast- queue)

Output geocast (m) $_{obj,i}$

Precondition:

Peek (geocast-queue) =m

Effect:

Dequeue (geocost- queue)

Input get-update (l,t) $_{obj,i}$

Effect:

Location ⬚ l

Clock⬚ t

Fig. 8 FPE server transitions for client i and object obj of variable type τ = <V, v$_0$, invocations, responses, δ >.

III. A SPECIFICATION IN TUG

TUG specification language consists of 3 parts: a name part where the title with input/ output parameters is given, on analysis part where the input data is defined, and an anatomy part where the output data is generated. The name part contains a module or schema title with input/ output parameters are enclosed in parentheses. The analysis part contains the rules for analyzing the input data. To analyze the input data, Definite Clause Grammars (DCG$_S$) are used to represent the rules to perform the syntax analysis. Each rule of a DCG expresses a possible form for a non terminal, as a sequence of terminals with optional constraints on the terminals and non terminals. Non terminal nodes in uppercase indicate constituents. A terminal node in lowercase indicates a

taken that must occur in the input data. A terminal node can be a literal which is any string enclosed in a pair of quotes. A constraint wrapped in braces places the conditions such as type checking indicates a taken that must occur in the input data. A terminal node can be a literal which is any string enclosed in a pair of quotes. A constraint wrapped in braces places the conditions such as type checking on a terminal node. Table I includes all operators used in the conditions. An input is parsed into a tree representation that takes the form of a prolog list with a node name acting as the relationship symbol of the input data [4]. This tree representation will be the input to the anatomy analysis part of the TUG specification.

Operators	Description
any (t)	t belongs to any type
bool (t)	t is a Boolean
character (t)	t is a character
digit (t)	t is a digit
equal_ to (t_1, t_2)	t_1 is equal to t_2
float (t)	t is a float
greater_ than (t_1, t_2)	t_1 is greater than t_2
integer (t)	t is an integer
length (t)	length of a string t
less_than (t_1, t_2)	t_1 is less than t_2
lowercase_charater (t)	t is a lowercase character
member (t_1, t_2)	t_1 is a member of t_2
not (t)	logical negation of t
remainder (t_1, t_2)	remainder for integer division t_1 / t_2
string (t)	t is a string
text (t)	t is a text
uppercase_ character (t)	t is an uppercase character
word (t)	t is a word

TABLE I OPERATORS USED IN THE CONDITIONS

The specification is structured with regular expression notations (Union, Positive Closure, and Concatenation). Each non terminal node structures its tree according to one of the regular expression notations. Adding tree structures to a specification allows a software developer to construct the specification in a structured way for dealing with complexity [4].

The union notation is indicated by a vertical bar sign (|) suffixed to the node name, so that

FLAG |

"On"

"Off"

Indicates that FLAG is one of the alternatives "on", and" off".

The concatenation notation is indicated by an ampersand sign (**&**) suffixed to the node name, so that

NAME &

Last_name

{String (last_name)}

First _ name

{String (first_ name)}

Indicates that NAME is a concatenation of last_ name and first_ name. The last_name and first_name must be of string type.

The kleene closure notation (*****) means zero or more element, over the node. Thus,

MAIL- Box*

Mail

{Text (mail)}

Indicates that NAIL- Box contains zero or more mails, each of which is a text.

The positive closure notation (**+**) means one or more elements over the node. So, that:

MAIL+
Message *
Letters
{Letter (letters)}

Indicates that MAIL contains at least one message, which may contain zero or more letters in it. If a message contains no letters, then the MAIL is an empty Mail

IV. RAPID PROTOTYPING PROCESS VIA SOFTWARE TRANSFORMATIONS

The prototype serves as a basis for discussion to help the specifier and the user to read just the user requirements and specifications. Feedback from the user is obtained to decide whether the change is minor or major. If the change is minor, A Change Request Script (CRS) specifying the change is written to update the specification and the prototype. If a major change is needed, the specifier may rewrite the specification and rederive a new prototype from the start. A major change may involve the structure of the specification to be modified. This prototyping process continues until the requirements have been thoroughly exercised and the user is satisfied with the demonstrated behavior of the prototype. The results of the prototype evolution are a set of modular TUG specifications for the proposed system. In addition, a set of CRSs record the design decisions made during the transformations [4].There is no existing specification language that can support abstractions for all features of Software systems. Therefore, a specification language must make developers to easily extend and modify prototypes. To support easy modifications to the prototype, the TUG specification language was designed to support the construction of a specification in a structured manner with regular expression notations. The modules in a derived prototype from the specification can be easily located, modified, or extended in terms of these regular expression notations. The rapid prototyping approach using TUG can be incorporated into any Software development process. It is intended that each evolution of the specification that is synthesized by the specifier should be formally recorded using TUG, and that the

prototype derived from the specification should be exercised with the users participating in the user requirements analysis process. The specification can then be reasoned with and expected behavior can be validated. The benefits of rapid prototyping have been identified to include [5]:-

- Rapid prototyping is available in the front end of the Software life cycle to allow early detection of errors.
- Unclear and imprecise user requirements can be clarified by rapid prototyping.
- Execution of the prototype supplements inspection and formal reasoning as means of analysis of the specification.
- The underlying theory of the TUG specification language is DCGs, which can be executed directly in the prolog environment .There is a close correspondence between TUG and prolog, which makes the process of transformation relatively mechanical. In this approach, DCGs are used as an intermediate form for aiding the transformation process. Although DCGs are syntactic for prolog, a prototype in DCGs seams difficult to read, understand, and maintain.
- Whenever there is a change in the user requirements, there many are no need to rederive the prototype from scratch if the change is trivial. ACRS is written to update the prototype only in response to the revised specification .A redervation of prototype in prolog from the start is avoided   is modified [6] [7].
- The rapid prototyping approach supports formal requirement specifications written in TUG.
- The prototype is exercised to demonstrate the system behavior in the prolog environment. A driver that reads in the input data and then calls the main program with parameters needs to be constructed manually. The set of transformation rules are given below. The conventions are:

- C: is a finite set of condition tests and has the for

$\{C1, C2,….,Cn\}$ with n ≥ 1 where c_i is a TUG condition test;

- Y is a finite set of dummy non terminal or terminal node;

- Names of predicates in prolog are in all lowers case letters.

- Names of variables in prolog are in all upper case letters.

- Q is a finite set of prolog procedure calls and has the form

$\{q_1, q2…q_n\}$ with n ≥ 1 where q_i is a prolog predicate for which a Prolog definition has been given, and. $<>$ encloses optional syntactic Items [8].

** The following four rules translate the analysis part of the TUG specification into DCGs. Each non-terminal node in the analysis tree structures its subtrees according to one of the structure notations each structuring operation can be transformed into a DCG form by applying the following four rules in straight forward manners[5][9].

$$|\text{-def 1}$$
$$\alpha 1$$
$$\beta 1 \text{ where } \alpha \text{ is a nonterminal node}$$
$$< \{\Phi 1\}>$$
$$\beta_2$$
$$< \{ \Phi_2\}>$$
$$…………..$$

$$…………..$$
$$\beta_n$$
$$<\{\Phi_n\}>$$
$$\alpha \to \beta_1 < \{ \Phi_1\}>$$
$$\alpha \to \beta_2 < \{ \Phi_2\}>$$
$$……………$$
$$……………$$
$$\alpha \to \beta_n < \{ \Phi_n\}>$$

In rule1, a union nonterminal node α in the analysis tree indicates that α is one of the alternatives, $\beta 1$, $\beta 2,…..$, β_n. If β_i is a literal, there is no Φ_i associated with β_i. Each translated DCG represents an alternative [9] [10].

$$\text{\&- def 2}$$
$$\alpha \, \&$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$< \{\Phi 1\}> \quad \beta_i \text{ is a nonterminal}$$
$$\text{or terminal node with condition tests } \Phi_i \quad C$$
$$\beta_2$$
$$<\{\Phi_2\}>$$
$$-----$$
$$-----$$
$$\beta_n$$
$$<\{ \Phi_n\}>$$
$$\alpha \to \beta_1 <\{ \Phi_1\}> \beta_2 <\{ \Phi_2\}> \quad …….. \quad \beta_n <\{ \Phi_n\}>$$

In rule 2, α a concatenation nonterminal node α in the analysis tree indicates that & is a concatenation of β_1, $\beta_2,…..$, β_n. If β_i is a literal, there is no Φ_i associated with β_i. Each translated DCG represents a concatenation form

$$\text{*def 3}$$
$$\alpha \, *$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$< \{\Phi_1\}> \quad \beta_1 \text{ is a nonterminal}$$
$$\text{or terminal node with condition tests } \Phi_i \quad C$$
$$\beta_2$$
$$< \{ \Phi_2\}>$$
$$-------$$
$$-------$$
$$\beta_n$$
$$<\{\Phi_n\}>$$
$$\alpha \to [\quad]$$
$$\alpha \to \beta_1 <\{ \Phi_1\}> \beta_2 <\{ \Phi_2\}> . ….. ……….. \beta_n <\{ \Phi_n\}> \alpha$$

In rule 3, a kleene closure nonterminal node α in the analysis tree indicates that α is a sequence of zero or more occurrence of β_1, $\beta_2,…..$, β_n. If βi is a literal, there is no Φ associated with β_i two translated DCGs represent a kleene closure form [5][11].

$$+ - \text{def 4}$$
$$\alpha +$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$<\{ \Phi_1\}> \quad \beta_i \text{ a nonterminal or terminal node with condition}$$
$$\text{tests } \Phi_i \quad C$$
$$\beta_2$$
$$< \{ \Phi_2\}>$$
$$------$$
$$------$$
$$\beta_n$$

$$<\{\Phi_n\}>$$
$$\alpha \to \beta_1<\{\Phi_1\}> \beta_2<\{\Phi_2\}> \quad \ldots\ldots \quad \beta_n <\{\Phi_n\}>$$
$$\alpha \to \beta_1<\{\Phi_1\}> \beta_2<\{\Phi_2\}> \quad \ldots\ldots \quad \beta_n <\{\Phi_n\}> \alpha$$

In rule 4, α a positive closure nonterminal node α in the analysis tree indicates that α is a sequence of one or more occurrences of $\beta_1, \beta_2 \ldots \beta_n$. If β_i is a literal, there is no Φ_i

(1) SEQUENCE→UNSORTED- ID_S
(2) SEQUENCE →SORTED- ID_S
(3) UNSORTED→ List 1 +X{integer(X)}
        List 2+Y {integer(y)},
        Greater-than(X, Y)}
        Rest_of_elements
(4) SORTED →ASCENDING-SEQUENCE
(5) ASCENDING_SEQUENCE→ Element {integer (element)}

(6) ASEDING_SEQENCE→ Element {integer (element)}
        ASCENDING-SEQUENCE

The following four rules translate the anatomy tree of the specification into DCGs. The rules are similar to the rules for translating the analysis tree. The difference is that we use the **":- "** symbol instead of symbol **" →"**. The use of the **" →"** symbol in the rules for the analysis part pf a TUG specification denotes a derivation of a tree, an involvement of pattern matching, and an unification of variables with the input values in prolog and the use of the **" :-"** symbol in the rules for the anatomy part of the specification perfumes the same operations. The uses of the **" →"** and **":-"** symbols are just for the syntactic purpose. The out puts of the rules for the analysis part of a TUG specification produce a tree unified with the input values that is the input to the rules for the anatomy part of the specification. The rules for the anatomy part of a TUG specification reads in the tree and performs exact unifications on the variables to produce outputs

Another difference is that dummy nodes appear in the rules. The reason for having dummy nodes is that often only the parts of tree are referenced in the anatomy tree of the specification. The remaining unreferenced parts of the tree still need to be unified in the course of patter matching. Dummy nodes are obtained by referring back to the analysis tree of the specification [7] [12].

$$| - def \; 5$$
$$\alpha \;|$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$\beta_2 \quad \beta_i \text{ is a nonterminal node}$$
$$\ldots\ldots\ldots$$
$$\beta_n$$
$$\varepsilon: -Y_1 \text{ where } \varepsilon \text{ is uppercase using } \alpha$$
$$\varepsilon: -Y_2 \text{ where } Y_i \text{ is uppercase using } \beta_i$$
$$\ldots\ldots\ldots\ldots$$
$$\varepsilon: -Y_n$$

In Rule 5, a union nonterminal node α in the anatomy tree indicates that α is one of the alternatives, $\beta_1, \beta_2, \ldots, \beta_n$. Each translated DCG represents an alternative.
$$\& - def \; 6$$

associated with β_i. Two translated DCGs represent a positive closure form [5][12].

To demonstrate the use of transformation rules presented in this section, the application of Rules 1-4 to the analysis tree of the related work approach specification produces the following results

$$\alpha \;\&$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$\beta_2 \quad \beta_i \text{ is a nonterminal node or statement}$$
$$\ldots\ldots\ldots$$
$$\beta_n$$
$$\varepsilon:-Y_1<\{\Psi_1\}> Y_2<\{\Psi_2\}>\ldots\ldots Y_n<\{\Psi_n\}> \quad \text{Where } \Psi_1 \underline{C} \; Y$$
$$\varepsilon \text{ is uppercase using } \alpha$$
$$Y_i \text{ is uppercase using } \beta_i \text{ if } \beta_i \text{ is a nontreminal node;}$$
$$\text{Otherwise } Y_i = \beta_i$$

In Rule 6, a concatenation nonterminal node α in the anatomy tree indicates that α is a concatenation of $\beta_1, \beta_2, \ldots, \beta_n$. The translated DCG represents a concatenation form.

$$* \text{-def } 7$$
$$\alpha \;*$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$\beta_2 \quad \beta_i \text{ is a nonterminal node ,terminal node, or statement}$$
$$\ldots\ldots\ldots.$$
$$\beta_n$$
$$\varepsilon:- [\;\;]$$
$$\varepsilon:- Y_1<\{\Psi_1\}> Y_2\{\Psi_2\}>\ldots\ldots.Y_n<\{\Psi_n\}> \varepsilon \quad \text{where } \Psi_1 \underline{C} \; Y$$
$$\varepsilon \text{ is uppercase using } \alpha$$
$$Y_i \text{ is uppercase using } \beta_i \text{ if } \beta_i \text{ is a nontreminal node;}$$
$$\text{Otherwise } Y_i = \beta_i$$

In Rule7, a kleene closure nonterminal node α in the anatomy tree indicates that α is a sequence of zero or more occurrence of $\beta_1, \beta_2, \ldots, \beta_n$. Two translated DCG_s represent a kleene closure form.

$$+ \text{-def } 8$$
$$\alpha +$$
$$\beta_1 \quad \text{where } \alpha \text{ is a nonterminal node}$$
$$\beta_2 \quad \beta_i \text{ is a nonterminal node ,terminal node, or statement}$$
$$\ldots\ldots$$
$$\beta_n$$
$$\varepsilon:- Y_1<\{\Psi_1\}> Y_2\{\Psi_2\}>\ldots\ldots.Y_n<\{\Psi_n\}>$$
$$\varepsilon:- Y_1<\{\Psi_1\}> Y_2\{\Psi_2\}>\ldots\ldots.Y_n<\{\Psi_n\}> \varepsilon$$
$$\text{Where } \Psi_1 \underline{C} \; Y$$
$$\varepsilon \text{ is uppercase using } \alpha$$
$$Y_i \text{ is uppercase using } \beta_i \text{ if } \beta_i \text{ is a nontreminal node;}$$
$$\text{Otherwise } Y_i = \beta_i$$

In Rule 8, a positive closure nonterminal node α in the anatomy tree indicates that α_{is} a sequence of one or more occurrences of $\beta_1, \beta_2, \ldots, \beta_n$.
Two translated DCG represent a positive closure form. The application of Rules 5-8 to the anatomy tree of the problem specification produces the following results

(7) SEQUENCE:- UNSORTED
(8) SEQENCE:-SORTED
(9)UNSORTED: -T-L1=Y:: List 2
   T-L2=X:: rest_of_elements
   T-L=List 1<>T-L1<>T-L2
   Call ID_S_sort (T-L)

(10) SORTED: - ASCENDING_SEQUENCE
(11) ASCENDING_SEQUENCE:-Output element
Output ' '
(12) ASCEDING_SEQENCE:-Output element
Output ' '
ASSCENDING_SEQUENCE

V. A TUG SPECIFICATION FOR IMPLEMENTING ATOMIC READ/WRITE SHARED MEMORY IN MOBILE AD HOC NETWORKS APPLICATION

This section will illustrates the usage of TUG for implementing atomic read/ write shared memory in mobile ad hoc networks. A specification in TUG is formalized incrementally in a modular and Top-down manner the example also illustrates how the language supports module independence via the language patterns. The Geoquorums approach, for implementing atomic read/ write shared memory in mobile ad hoc networks. This approach is based an associating abstract atomic object, with certain geographic locations. We assume the existence of local points, geographic areas that are normally "populated" by mobile nodes. The Geoquorum algorithm uses the fault –prone focal point objects to implement atomic read/write operations on fault – tolerant virtual shared object. Te Geoquorums algorithm uses a quorum- based strategy in which each quorum consists of a set of focal point objects. The quorums are used to maintain the consistency of the shared memory and to tolerate limited failures of the focal point objects which may be caused by depopulation of the corresponding geographic areas .Overall, the new geoquorums algorithm efficiently implements read and write operations in a highly dynamic, Mobil network.

A. A First Attempt at the Specification

MODULE a_ listing_ of _transitions
(in: TRANSITION_TYPE)
ANALYSIS
TRANSITION _TYPE &
TRANSITIONS *
"Put"
"get"
"confirm"
"config"
"reconfig"
END OF ANALYSIS"
ANATOMY
Transition_ type&
Transition*
Output n1
Output "put"
Output "get"
Output "confirm"

Output "config"

Output "reconfig"

END OF ANATOMY

END OF MODULE a_ Listing_ of_ transitions

B. The Application of Transformation Rules to the Above Specification Module Results

In the Following Prototype in Prolog:

Transition_ type (transition_ type

(TRANSITION))→

Transition (TRANSITION).

Transition (Transition ([ ]))→

[ ].

Transition (transition ("put", "get", "confirm", "config", "reconfig")) →

["put"],

["get"],

["confirm'],

["config"],

["reconfig"]

Transition (TRANSITION).

Transition_ type (transition_ type

((TRANSITION)):-

Transition (REANSITION).

Transition (Transition ([ ])).

Transition (Transition("put", "get", "confirm", "config", "reconfig")) :-

n1,

Write ("put", "get", "confirm", "config", "reconfig")),

Transition (TRANSITION).

(2)The Following CRS is Further Refinement on Each Invocation

Replace TRANSITION* under TRANSITION _Type &

With

VARIABLE_TYPE_TRANSITION.

"put_ invocation"

"get_ invocation"

"confirm_ invocation"

"config _invocation"

"reconfig_ invocation"

Replace Transition* under transition_ type &

With

Variable_Type_Transition |

Variable Transition &

Output n1

Output "put _ invocation"

Output" get _ invocation"

Output "confirm – invocation"

Output "config – invocation"

Output" reconfig – invocation"

C. A CRS for This Refinement is shown below

Replace PUT_INVOCATION_SECTIONS* under PUT_ INVOCATION

With

Put_ invocation

{new_value (put_invocation)

New_tag (put_ invocation)

New_config_id (put_ invocation)}

"get- invocation".

ReplaceGET_INVOCATION_SECTIONS*underGET_ INVOCATION & with

get_invocation

{new_config_id (get_invocation)}

"get_invocation"

At this stage, the application of transformation rules to the above two CRSs result in the following Prolog to update the prototype.

Transition_Type (transition_type (TRANSITION)) →

transition (TRANSITION).

transition (transition ([ ])) → [ ]

transition(transition(PUT_INVOCATION,GET_INVOCATI ON, CONFIRM_INVOCATION,CONFIG_INVOCATION,RECO NFIG_INVOCATION)) →

put_invocation (PUT_INVOCATION),

(GET_INVOCATION).

transition (TRANSITION)

Variable_type_transition(put_invocation (PUT_INVOCATION)) →

put_invocation (PUT_INVOCATION).

Variable_type_transition(get_invocation (GET_INVOCATION)) →

get_invocation (GET_INVOCATION).

put_invocation('put_invocation_separator' PUT_INVOCATION_SECTION)) →

['put_invocation_separator'],

put_invocation_section (PUT_INVOCATION_SECTION).

put_invocation_section ([ ])).

put_invocation_section(put_invocation_section(PUT_INVOC ATION,'put_invocation_separator', PUT_INVOCATION_SECTION)) →

[PUT_INVOCATION],

{put_ack_response (PUT_INVOCATIO, NEW_VALUE, NEW_TAG,

NEW_CONFEG_ID),

Length (new_config_id >config_id)

[put_invocation→ PUT_ACK_RESPONSE],

Put_invocation_section (PUT_INVOCATION_SECTION).

Transition_type (transition_type (TRANSITION)): -

transition (TRANSITION)

transition (transition ([ ])).

Transition(transition(PUT_INVOCATION, GET_INVOCATION,CONFIG_INVOCATION, RECONFIG_DONE _INVOCATION)): -

get_invocation (GET_INVOCATION),

transition (TRANSITION).

Get_invocation (get_invocation (GET_INVOCATION)):-

get_invocation (GET_INVOCATION).

confirm_invocation (confirm_ invocation (CONFIRM- INVOCATION)):

confirm_ invocation (CONFIRM_ INVOCATION).

get_invocation(get_invocation ('initial_get_invocation_separator',

GET_INVOCATION_SECTION)):-

write ('get_invocation'),

write ('put_invocation'),

write ('config_invocation'),

write ('reconfig _done_invocation'),

confirm_invocation(confirm_invocation('new_tag',

CONFIRM_INVOCATION_SECTION)): -

nl,

write ('put_invocation'),

write ('get_invocation'),

write ('config_invocation'),

write ('recon_done_invocation'),

D. The Further Refinement

Replace 'put_invocation' under

NEW_CONFIG_ID & with

PUT_ACK_RESPONSE

Stop &

'Stop'

' '

replace 'get_invocation' under

NEW_CONFIG_ID & with

GET_ACK_RESPONSE

Stop&

'Stop'

' '

replace 'config_invocation' under

NEW_TAG & with

CONFIG_ACK_RESPONSE

Stop&

'Stop'

' '

replace 'recon_done_invocation' under

NEW_CONFIG_ID & with

RECON_DONE _ ACK

Stop&

'Stop'

' '

replace transition_ type & with

output 'transition Analysis'

output nl

transition*

output 'transition:'

output 'put_invocation'

output 'put _ ack _ response'

output 'get_invocation'

output 'get _ ack _ response'

output '**.**'

output nl

config_invocation &

output ' config_invocation'

output ' config _ ack _ response'

output '**,**'

output ' recon_done _ invocation'

output ' recon_done _ ack'

output '**.**'

output n1

After a successive of refinements to the original specification, the final complete specification for implementing atomic read/write shared memory in mobile ad hoc networks is shown below.

MODULE a _ Listing _ of _ Transitions

(in: TRANSITION _ TYPE)

ANALYSIS

TRANSITION _ TYPE &

TRANSITION*

PUT_INVOCATION |

NEW_CONFIG_ID&

PUT_ACK_RESPONSE

Stop&

'Stop'

Get_ INVOCATION |

NEW_CONFIG_ID&

GET_ACK_RESPONSE

Stop&

'Stop'

' '

CONFIG_INVOCATION |

NEW_TAG&

CONFIG_ACK_RESPONSE

Stop&

'Stop'

' '

RECONFIG_DONE_ INVOCATION |

NEW_CONFIG_ID&

RECON_DONE_ACK

Stop&

'Stop'

' '

END OF ANALYSIS

ANATOMY

transition_type &

output' Transition Analysis'

output nl

transition*

output' transition:'

put _invocation |

new_config_id &

output 'put_ invocation'

output 'put_ack_response'

output '. '

output nl

get _ invocation |

new_config_id &

output 'get_ invocation'

output 'get_ack_response'

output ', '

config- invocation |

new-tag &

output 'config- invocation'

output 'config -ack-response'

output ' , '

recon - done- invocation |

new-config –id &

output ' reconfig - done- invocation'

output ' recon - done- ack'

output '.'

output nl

END OF ANATOMY;

END OF MOUDULE a-Listing-Of-Transitions.

VI. CONCLUSIONS AND FUTURE WORK

An approach was developed to support rapid prototyping via software transformations by deriving a prototype in prolog from a specification in TUG. We didn't directly use the prolog language to specify user requirements, programming languages more or less concentrate on how rather than what and are generally unsuitable for specification purposes. In addition, in a specification in prolog lacks modularity in contrast to a specification in TUG. Since the main purpose of a specification is to aid the understanding of the user requirements, it is useful if a specification can be read and understood. Modularity helps specifiers to read and maintain in a manageable way. TUG provides modularity to help specifiers to specify a system in a hierarchical manner. A set of modules are specified and then composed into a system. The system is tested in pieces corresponding to the modular specification .In contrast to a specification in TUG, a specification in prolog is relatively difficult to maintain. Rapid prototyping via software transformations helps to build prototypes automatically from specifications. In this paper, a formal method with TUG was presented to support the rapid prototyping via software transformations process in which a prototype can be built quickly and cheaply. Automation of the application of software transformations reduces the labor intensity of developing prototypes manually. Rapid prototyping via software transformations also provides support for prototype modifications. The rapid prototyping approach supports prototype evolution by avoiding complete retransformation of the prototype from the start whenever there is a change made to the specification. To avoid complete retransformation, a CRS is written to update the prototype only in response to the minor changes to the specification, involving the nodes to be modified, extended, relaxed, or refined. If a major change is needed, the specification may need to be rewritten and a new prototype may be derived from the start. A major change may involve the structure of the specification to be modified. Like other formal specification languages, the TUG specification language may not provide enough abstractions for modeling some properties of software systems such as non- functional properties. Therefore, the geoquorum approach for implementing atomic read/write shared memory in mobile ad hoc networks encourages specifiers to manually add additional code to the derived prototype for demonstrating such kind of properties of systems. The rapid prototype approach supports the quick construction of a prototype with a high degree of module independence. Module independence has a particular importance in this approach because of the need for modifications to the prototype. It remains an open question to determine how to choose a good set of focal points, how to construct a map of focal points in a distributed fashion, and how to moodily the set of focal points dynamically. Overall, the FPO Model will significantly simplify the development of algorithms for mobile, in highly dynamic networks. Finally, there exist many techniques to do the phases of software lifecycle for any application.

REFERANCES

[1] C.Chiang,J.E.Urban, "Validating Software Specification against User Claims", Proceedings of the Twenty-Third Annual International Computer Software and Applications Conference (COMPSAC 2000),2000,PP:104-109.

[2] DOLEV, S., Gilbert, S.LYNCH, N.A., SHVARTSMAN, A.A., Welch, J.L.: " Geoquorums: Implementing Atomic Memory in Mobile Ad Hoc Networks". In: Proceeding of the 17th International Conference on Distributed Computing, PP. 306-320 (2003).

[3] Haas, Z.J., Liang, B.: "Ad Hoc Mobile Management with Uniform Quorum Systems". IEEE/ACM Transactions on Networking 7(2), PP: 228-240 (2000).

[4] B.CMoszkowski,"A Complete Axiomatization Of Interval Temporal Logic With Infinite Time ", Proceedings of the 15th Annual IEEE Symposium on Logic in Computer Science (LICS·00),JUNE(2000),26-29,California ,2000,PP:242-249.

[5] Chia- Chu Chiang," Automated Rapid Prototyping Of Tug Specifications Using Prolog ", Proceedings Of: Information And Software Technology 46(2004), PP: 857-873.

[6] Chia-Chu Chiang," Automated Rapid Prototyping of TUG Specifications Using Prolog", Proceedings of: Information and Software Technology 46(2004), PP: 857-873.

[7] O.J.Dahl, O. Owe, Formal Methods and the RMODP, Research Report No.261, Department of Information, University of Oslo, Norway, 1998.

[8] A. Evans, UML Class Diagrams –Filling the Semantic Gap, Technical Report, York University, 1998.

[9] IEEE, IEEE standard for a High Performance Serial Bus, Standard 1394, August 1995.

[10] M.Liu Yanguo, Proof Patterns for UMI-Based Verification, Master Thesis ECE Department, University of Victoria, Victoria, Canada, October 2002.

[11] I.Traore, D.Aredo, H.Ye," An Integrated Framework for Formal Development of Open Distributed Systems," In: Information and Software Technology 46 (2004) 281-286.

[12] I. El-Far, Automated Construction of Software Behavior Models, Master's Thesis, Florida Institute of Technology Melbourne, Fl, 1999.

15

PSS Design Based on RNN and the MFA\FEP Control Strategy

Rebiha Metidji and Boubekeur Mendil
Electronic Engineering Department
University of A. Mira, Targua Ouzemour,
Bejaia, 06000, Algeria.
Zeblah80@yahoo.fr

Abstract –   **The conventional design of PSS (power system stabilizers) was carried out using a linearized model around the nominal operating point of the plant, which is naturally nonlinear. This limits the PSS performance and robustness.**

In this paper, we propose a new design using RNN (recurrent neural networks) and the model free approach (MFA) based on the FEP (feed-forward error propagation) training algorithm [15].

The results show the effectiveness of the proposed approach. The system response is less oscillatory with a shorter transient time. The study was extended to faulty power plants.

Keywords-Power Network; Synchronous Generator; Neural Network; Power System Stabilizer; MFA/FEP control.

I. INTODUCTION

Power system stabilizers are used to generate supplementary control signals for the excitation system, in order to damp the low frequency power system oscillations. Conventional power system stabilizers are widely used in existing power systems and have made a contribution in enhancing power system dynamic stability [1]. The parameters of a classical PSS (CPSS) are determined based on a linearized model of the power system around its nominal operating point. Since power systems are highly nonlinear with time varying configurations and parameters, the CPSS design cannot guarantee good performance in many practical situations.

To improve the performance of CPSSs, numerous techniques have been proposed for their design, such as the intelligent optimization methods [2], fuzzy logic [3,4,5], neural networks [7,8,9] and many other nonlinear control techniques [11,12]. The fuzzy reasoning using qualitative data and empirical information make the fuzzy PSS (FPSS) less optimizing compared with the neural PSS (NPSS) performance. This is our motivation.

The main problem in control systems is to design a controller that can provide the appropriate control signal to meet some specifications constituting the subject of the control action. Often, these specifications are expressed in terms of speed, accuracy and stability. In the case of neuronal control, the problem is to find a better way to adjust the weights of the network. The main difficulty is how to use the system output error to change the controller parameters, since the physical plant is interposed between the controller output and the scored output.

Several learning strategies have been proposed to overcome this problem such as the supervised learning, the learning generalized inverse modeling, the direct modeling based on specialized learning, and so on [14]. In this work, we used our MFA/FEP approach because of its simplicity and efficiency [15]. The aim is to ensure a good damping of the power network transport oscillations. This can be done by providing an adequate control signal that affects the reference input of the AVR (automatic voltage regulator). The stabilization signal is developed from the rotor speed or electric power variations.

The section II presents the power plant model. The design of the neural PSS is described in section III. Some simulation results are provided in section IV.

II. THE POWER PLANT MODEL

The standard model of a power plant consists of a synchronous generator, turbine, a governor, an excitation system and a transmission line connected to an infinite network (Fig.1). The model is built in MATLAB/SIMULINK environment using the power system Blockset. In Fig.1, P_{REF} is the mechanical power reference, P_{SR} is the feedback through the governor, T_M is the turbine output

torque, V_{inf} is the infinite bus voltage, V_{TREF} is terminal voltage reference, V_t is terminal voltage, V_A is the voltage regulator output, V_F is field voltage, V_E is the excitation system stabilizing signal, Δw is the speed deviation, V_{PSS} is the PSS output signal, P is the active power, and Q is the reactive power at the generator terminal.

The switch S is used to carry out tests on the power system with NPSS, CPSS and without PSS (with switch S in position 1, 2, and 3, respectively).

The synchronous generator is described by a seventh order d-q axis of equations with the machine current, speed and rotor angle as the state variables. The turbine is used to drive the generator and the governor is used to control the speed and the real power. The excitation system for the generator is shown in Fig.2 [4].

The CPSS consists of two phase-lead compensation blocks, a signal washout block, and a gain block. The input signal is the rotor speed deviation $\Delta\omega$ [16]. The block diagram of the CPSS is shown in Fig.3.
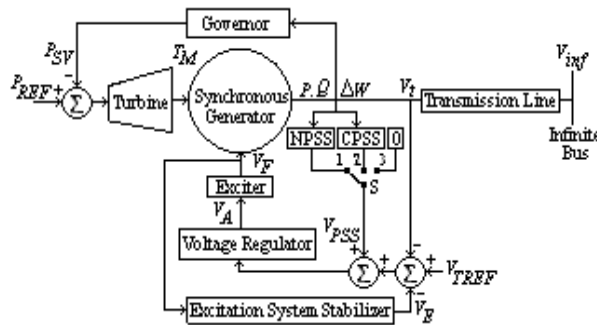
Figure 1. The control system configuration.
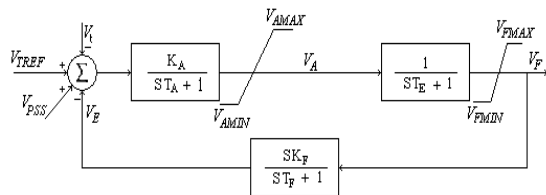
Figure 2. Block diagram of the excitation system.
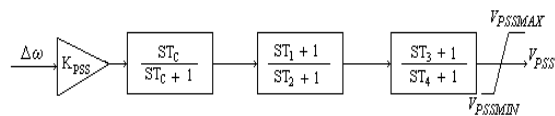
Figure 3. Block diagram of CPSS.
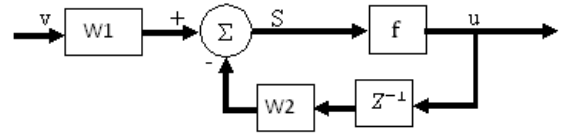
Figure 4. Block diagram of DTRNN.
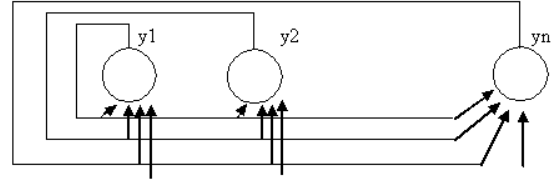
Figure 5. Illustration of the state feedback.

III. THE NPSS DESIGN

In this work, we used a neural architecture used primarily in the field of modeling and systems dynamic control; namely the DTRNN (Discrete Time Recurrent Neural Network). This is analogous to the Hopfield network (Fig.4).

W : global synaptic weight vector.

S : weighted sum, called potential.

U : output or neuron response.

f : activation function.

υ: external input.

This is a neural network with state feedback. It has a single layer network with the output of each neuron is feed back to all other neurons. Each neuron may receive an external input [14]. This is well shown in Fig.5.

A. DTRNN network Equations

Originally, the equations governing this network type are the form:

$$U_i(k+1) = f[\sum_{j=1}^{m} W_{ij} \times U_j(k) + \nu_i(k)] \quad (1)$$

If we consider that the inputs ν_i are weighted as in Fig.4, (1) can be rewritten as follows:

$$U_i(k+1) = f[\sum_{j=0}^{m+n} W_{ij} \times U_j(k)] \quad (2)$$

With

$$U_j(k) = \begin{cases} 1 & j = 0 \\ U_j(k) & j = 1, \dots, n \\ \nu_{j-n}(k) & j = n+1, \dots, n+m \end{cases} \quad (3)$$

Where:

$U_i(k+1)$: i^{th} Network state variable (i=1… n).

$\nu_j(k)$: j^{th} External input (j=1… m).

n : number of neurons in the network.

m : dimension of the input vector, V.

$U_{0i}(k)=1$: inner threshold.

The outputs are chosen among the state variables. Learning is done using dynamic training algorithms that adjust the synaptic coefficients based on presented examples (corresponding desired inputs \ outputs) and taking into account the feedback information flow [17].

In this work, we used the MFA training structure of Fig.6. The idea is to evaluate the output error (i.e. the gap between the system output and its desired value) to the controller input and to spread them directly from input to output, to get the hidden layers and the output layer errors. This can be realized by the feed forward error propagation algorithm (FEP) crafted specifically for this purpose. This allows direct and fast errors calculation of consecutive layers, required for the controller parameters adjustment.

B. Training the DTRNN controller based on the algorithm FEP

The approach MFA / FEP is an effective alternative for training controllers. Direct injection of the input error can provide error vector components required to the network weight update. The FEP formulation is given in the following paragraph.

Let consider a DTRNN given by (2) and (3). The global input vector, X (k), of the network consists of the threshold input, d=1, the external inputs, xi (k), and network state feedback variables, U_i (k).

$$X(k) = \begin{bmatrix} 1 \\ X_1(k) \\ . \\ . \\ X_m(k) \\ u_1(k-1) \\ . \\ . \\ u_n(k-1) \end{bmatrix} = \begin{bmatrix} X_0(k) \\ X_1(k) \\ . \\ . \\ . \\ . \\ . \\ X_L(k) \end{bmatrix} \quad \text{and} \quad U(k) = \begin{bmatrix} U_1(k) \\ U_2(k) \\ . \\ . \\ . \\ U_n(k) \end{bmatrix}$$

Then, we can write

$$U_i(k+1) = f[S_i(k)] = f[\sum_{j=0}^{m+n} W_{ij}.X_j(k)] \quad (4)$$

The outputs are chosen among the state variables. For the simplicity of notation, let consider the first *r* state variables as outputs.
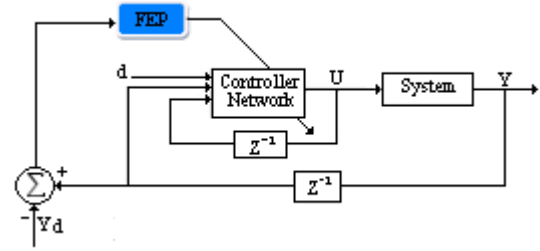
Figure 6. MFA/FEP control structure.

$$Y(k) = [y_1(k), y_2(k),…,y_r(k)]^T$$

$$= [U_1(k), U_2(k),… U_r(k)]^T \quad (5)$$

The standard quadratic error over the *pth* sequence is given by

$$J_p(W) = \frac{1}{2}\sum_{k=1}^{k_p} \sum_{j=1}^{r}[y_j(k) - y_j^d(k)]^2 \quad (6)$$

Where

k_p : is the length of the *pth* training sequence,

r : is the number of the network outputs,

$y_j(k)$: is the *jth* network output,

$y_j^d(k)$: is the desired value corresponding to $y_j(k)$.

The weights are updated iteratively according to

$$W_{ij}(k+1) = W_{ij}(k) - \mu \frac{\partial J_p(W)}{\partial W_{ij}(k)} \quad (7)$$

where μ is the learning rate. The gradient in (7) is calculated using the chain rule:

$$\frac{\partial J_p(W)}{\partial W_{ij}(k)} = \frac{\partial J_p(W)}{\partial U_i(k)} \frac{\partial U_i}{\partial W_{ij}(k)} \quad (8)$$

Where

$$\frac{\partial U_i}{\partial W_{ij}(k)} = \frac{\partial f(s_i(k))}{\partial s_i(k)} \frac{\partial s_i(k)}{\partial W_{ij}(k)}$$

$$= f'[s_i(k)].x_j(k) \quad (9)$$

The term $\frac{\partial J_p(W)}{\partial U_i(k)}$ in (8) is the sensitivity of Jp(w) to the node output, $U_i(k)$. Let the error, calculated between the network output and desired output, be

$$E_0(k) = [Y(k) - Y^d(k)] = \Delta Y(k) \quad (10)$$

As in the back-propagation algorithm, the term $\frac{\partial J_p(W)}{\partial U_i(k)}$ can be interpreted as the equivalent error, $\varepsilon_i(k)$, with (*i=1,2,...,n*). Hence, we write

$$\frac{\partial J_p(W)}{\partial U_i(k)} = \varepsilon_i(k) = \sum_{j=1}^{r} \frac{\partial f(s_i(k))}{\partial y_j(k)} \Delta y_j(k) \quad (11)$$

The error vector, $E_0(k)$, can be calculated at the input ; since the network receives the output vector, Y(k), as input through the state feedback. The

equivalent error vector components, $\varepsilon_i(k)$, are computed using (11), through the forward propagation of $E_0(k)$.

C. *NPSS design using the MFA / FEP approach*

The aim is to ensure an optimal control of power system output voltage. The MFA / FEP learning structure is illustrated by Fig.7.

The adequate structure of the NPSS network consists of 3 neurons. The input vector X= $[d=1, dw, w, U^T]^T$, where $U_{(3x1)}$ is its own state vector, d is the threshold input, *w* is the angular velocity, and *dw* is its variation. The NPSS output, V_{NPSS}, chosen as the first state variable, is applied to the automatic voltage regulator input. The learning rate is fixed at $\mu = 0.2$.

The training procedure is summarized by the followed steps:

Step1- Initialize the weights to small values in an interval [0, 0.1]. The synaptic weight matrix is of size 3×6 (3 outputs and 6 inputs).

Step 2- Initialization the state variables.

Step 3- Compute NPSS output.

Step 4- Application to the single-machine infinite bus (SMIB) process.

Step 5- Compute the output error and update the NPSS weights.

*Step 6- G*o to step 3 or stop if the learning is completed.
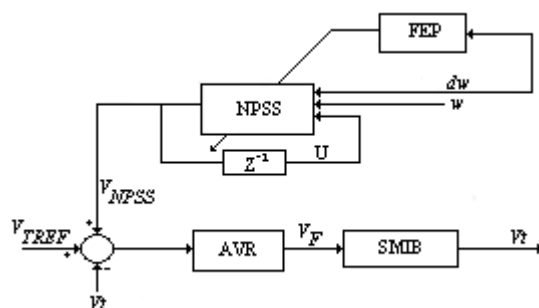
Figure 7. The NPSS training structure based on MFA / FEP approach.

IV. SIMULATION RESULTS

The parameters of the machine model used hang the simulation are:

-The nominal Power : $P_n = 3.125 \times 10^6$ (VA).

-The nominal voltage : $V_n = 2400$ (V).

-The nominal frequency : $f_n = 50$ (Hz).

To evaluate the performance of the NPSS, the system response of the NPSS is compared with the cases where there is no PSS and with a CPSS in the system.
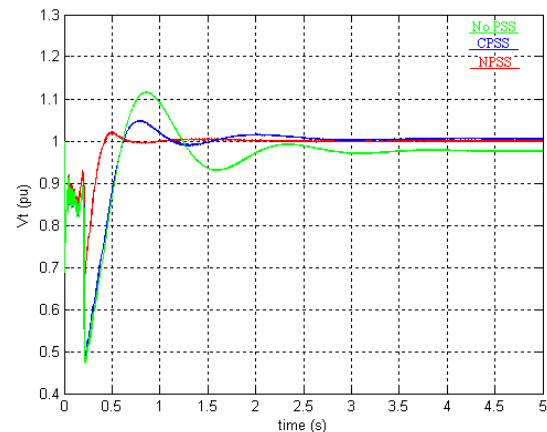
Figure 8. Terminal voltage response.
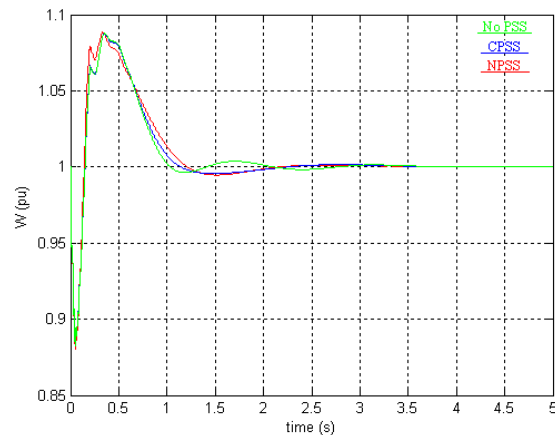
Figure 9. Machine angular speed.

From these results, we can see that the response of the system stabilizes at the reference value after 2 sec. But, the NPSS provide a better voltage response with reduced overshoot and rise time.

In order to evaluate the performance of the different PSS in the presence of defects, let us consider, as an example, a temporary short circuit, at t = 5s, that lasts 0.1s (time required for switching the circuit breaker protection).
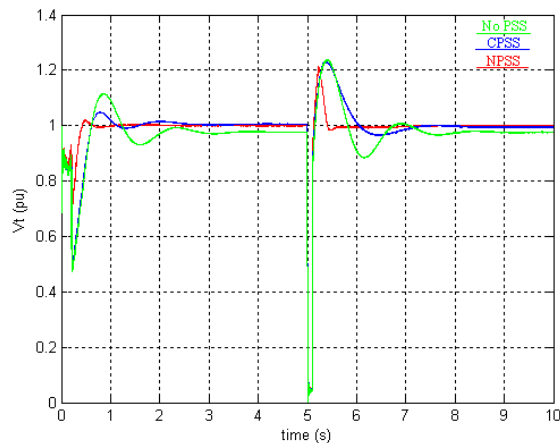
Figure 10. Terminal voltage response.
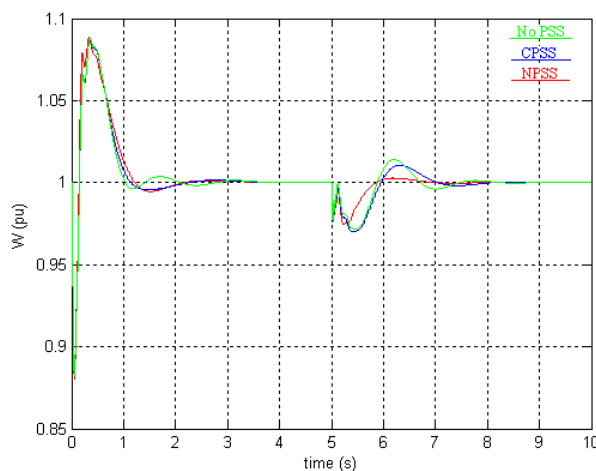
Figure 11. Machine angular speed.

The fault in the network has affected the output voltage and the angular speed. We also note that the system returns to steady state with less oscillations and with a shorter time using the control system (AVR) with the NPSS compared to CPSS and also with the system without PSS.

V. CONCLUSION

In this paper, we have developed another way to design PSS. The NPSS is a DTRNN trained using our MFA / FEP learning structure. The results show the effectiveness and the best performance of the resulting stabilizer.

REFERENCES

[1] S. M. Pérez, J. J. Mora, & G. Olguin, '' *Maintaining Voltage Profiles by using an Adaptive PSS*'', Proc. IEEE PES Transmission and Distribution Conf. and Exposition Latin America, Venezuela, 2006.

[2] L. H. Hassan, M. Moghavvemi, & H. A.F. Mohamed, ''*Power System Stabilization Based on Artificial Intelligent Techniques: A review*'' , International Conference for Technical Postgraduates (TECHPOS), pp. 1-6, 2009.

[3] A. Hariri & O.P. Malik, ''*A Fuzzy Logic Based Power System Stabilizer with Learning Ability*'', IEEE Trans on Energy Conversion, Vol. 11, No. 4, pp. 721-727, Dec 1996.

[4] M. Chetty, ''*A Fuzzy Logic Based Discrete Mode Power System Stabilizer*'', Asian Journal of Cont, Vol. 4, No. 3, pp. 327-332, Sep 2002.

[5] P. V. Etingov & N. I. Voropai,'' *Application of Fuzzy Logic PSS to Enhance Transient Stability in Large Power Systems*'', Proc. Interna. Conf. on Power Electronics, Drives and Energy Systems, pp. 1 - 9, 2006.

[6] A.A. Gharaveisi, S.M.R. Rafiei, & S.M. Barakati ''*An Optimal Takagi-Sugeno Fuzzy PSS For Multi- Machine Power System*'', Proc of the 40th North American Power Sympo, pp. 1-9, 2008.

[7] P. Shamsollahi & 0. P. Malik, ''*An Adaptive Power System Stabilizer Using On-Line Trained Neural Networks*'', IEEE Trans. on Energy Conversion, Vol. 12, No. 4, pp. 382 -287, Dec 1997.

[8] C.-J. Chen, T.-C. Chen, & J.-C. Ou, ''*Power System Stabilizer Using a New Recurrent Neural Network for Multi- Machine*'', Proc. IEEE Interna. Power and Energy Conf, pp. 68 – 72, 2006.

[9] C.-J. Chen, T.-C. Chen, H.-J. Ho & J.-C. Ou'' *PSS Design Using Adaptive Recurrent Neural Network Controller* '', Proc. Interna. Conf. on Natural Computation, 2009. ICNC '09. Fifth, Vol.2, pp. 277-281, 2009.

[10] P. Tulpule & A. Feliachi, ''*Online Learning Neural Network based PSS with Adaptive Training Parameters*'', IEEE Power Eng. Society General Meeting, pp. 1-5, 2007.

[11] S.-M. Baek & J.-W. Park, '' *Nonlinear Controller Optimization of a Power System Based on Reduced Multivariate Polynomial Model*'', Proc. Interna. Joint Conf. on Neural Networks, pp. 221 -228, 2009.

[12] H. Amano & T. Inoue, '' *A New PSS Parameter Design Using Nonlinear Stability Analysis*'', IEEE Power Engineering Society General Meeting, pp. 1- 8, 2007.

[13] S. Panda & C. Ardil, ''*Robust Coordinated Design of Multiple Power System Stabilizers Using Particle Swarm Optimization Technique*'', Interna. Journal of Electrical and Electronics Eng. 1:1 2008.

[14] J. He, ''*Adaptive Power System Stabilizer Based On Recurrent Neural Network* '', PhD thesis, Univ. of Calgary, 1998.

[15] B. Mendil & K. Benmahammed, ''*Model Free Approach for Learning Control Systems*'', Interna. Journal of Computer Research, Vol. 11, n° 2, pp 239-247, 2002.

[16] A. A. Gharaveisi, M. Kashki, S.M.A. Mohammadi, S.M.R. Rafiei, & S.M. Vaezi-Nejad, ''*A Novel Automatic Designing Technique for Tuning Power System Stabilizer* '', Proc. of the World Congress on Eng. Vol III, 2008.

[17] D. R. Hush & B.G. Horne*, ''Progress in Supervised Learning Neural Networks*'', IEEE signal proc. magazine, Vol.10, No.1, pp.8-39, Jan. 1993.

An Efficient SJRR CPU Scheduling Algorithm

Saeeda Bibi [1], Farooque Azam[1], Sameera Amjad[1], Wasi Haider Butt[1], Hina Gull[1], Rashid Ahmed[1], Yasir Chaudhry [2]

[1]Department of Computer Engineering
College of Electrical and Mechanical Engineering, NUST
Rawalpindi, Pakistan
[2]Department of Computer Science
Maharishi University of Management
Fairfield,Iowa USA
{saeedabb, sameerashaheen, butt.wasi, hinagull03, yasiryc }@gmail.com, farooq {farooq , rashid}@ceme.nust.edu.pk

Abstract— **CPU Scheduling is a vital discipline which helps us gain deep insight into the complex set of policies and mechanisms used to govern the order in which tasks are executed by the processor. This article proposes an efficient Shortest Job Round Robin (SJRR) CPU Scheduling algorithm having better average waiting time (AWT) and average turnaround time (ATT) as compared to other CPU Scheduling techniques. The primary objective of this algorithm is to optimize system performance according to the criteria deemed most important by system designers. Included in this work, is a simulation that compares the proposed algorithms with some well known practices to CPU scheduling.**

Keywords-component; First Come First Serve Algorithm, Shortest Job First Algorithm, Round Robin Algorithm, Priority Algorithm, Average Waiting Time, Turnaround Time, Response Time, Throughput

I. INTRODUCTION

Scheduling is a technique which involves complex set of policies and mechanisms working at the back of a processor, instructing it the order in which it should execute a given set of processes. Process is a smallest work unit of a program which requires a set of resources for its execution that are allotted to it by the CPU. These processes are many in number and keep coming in the queue one after the other. In order to execute them in a particular fashion, different scheduling techniques are employed that enable faster and efficient process execution thereby reducing the waiting time faced by each process and increasing CPU utilization. A process has five basic states namely New, Ready, Running, Waiting and Terminate [1] [5].

Throughout its lifetime a process migrates between various scheduling queues by different schedulers until it gets terminated. These queues mainly contain the ready queue which contains set of processes ready for CPU response. The second queue is the device or the I/O queue which contains all the processes that are waiting for I/O response [1]. The operating system must select processes for scheduling from these queues in a specific manner. This selection process using a particular scheduling technique is carried out by schedulers. Schedulers in general try to maximize the average performance of a system according to the given criterion [2].

Scheduling Algorithms can be broadly classified into preemptive and non-preemptive scheduling disciplines. The algorithm proposed in this article is preemptive in nature and attempts to give fair CPU execution time by focusing on average waiting time and turnaround time of a process. This article comprises of the following sections: Section 2 presents scheduling parameters which will decide against which parameters the new CPU scheduling algorithm will be tested. Section 3 introduces existing scheduling algorithm. Section 4 explains the SJRR scheduling algorithm. Section 5 contains pseudocode of the algorithm. Section 6 explains the two basic elements that make up the simulation and provide an interactive user interface. Section 7 presents a graphical comparison of the new algorithm with existing techniques. Last but not the least Section 8 will provide conclusion of the work...

II. SCHEDULING PARAMETERS

Different scheduling algorithms have different characteristics which decide selection of processes using different criteria for execution by CPU. The criteria which decide how one algorithm differs from the other have been listed below:

A. Processor utilization

It is the average fraction of time during which the processor is busy [2]. Being busy means the processor is not idle.

B. Throughput

It refers to the amount of work completed in a unit of time [2]. That is, the number of user jobs executed in a unit of time. The more the number of jobs, the more work is done by the system.

C. Turnaround Time

It is defined as the time taken to execute a given process [1]. That is, it is the time spends by the process in the system from the time of its submission until its completion by the system.

D. Waiting Time

Scheduling algorithms do not affect the amount of time during which a process executes or does I/O, it affects only the amount of time spend by the process in the ready queue [1]. That is, the amount of time spent in the ready queue by the process a waiting CPU execution.

E. Response Time:

While turnaround time includes total time taken by the process from the time of its submission until the time of its completion, response time is the measure of time from the submission of requests until the first response is produced [1]. This response time does not include the time taken to output that response.

III. OVERVIEW OF EXISTING CPU SCHEDULING ALGORITHMS

CPU scheduling algorithms aim at deciding which processes in the ready queue are to be allotted to the CPU. Discussed in this section are some common CPU scheduling algorithms

A. First Come First Served (FCFS) Scheduling

FCFS employs the simplest scheduling technique on the basis of first come first served. The work load is processed in the order of arrival, with no preemption [2]. Once a process has been submitted to the CPU, it runs into completion without being interrupted. Such a technique is fair in the case of smaller processes but is quite unfair for long an unimportant job [3]. Since FCFS does not involve context switching therefore it has minimal overhead. It has low throughput since long processes can keep processor occupied for a long time making small processes suffer. As a result waiting time, turnaround time and response time can be low [4].

B. Shortest Job First (SJF) Scheduling

Shortest Job First is non-preemptive in nature in which process with smallest estimated run time to completion is executed next. SJF reduces average waiting time of processes as compared to FCFS. SJF favors shorter processes over longer ones which is an overhead as compared to FCFS [6]. It selects the job with the smallest burst time ensuing CPU availability for other processes as soon as the current process reaches its completion.

This prevents smaller processes from suffering behind larger processes in the ready queue for a long time [3] [7].

C. Round Robin (RR) Scheduling

Round Robin is preemptive in nature. It employs FCFS for process execution by assigning a quantum or time slice to each process [7]. As soon as the quantum expires control is forcefully taken from the current process under execution and is transferred to the next in the queue for the same period of time slice [3]. The outcome of RR algorithm in term of performance depends entirely on the size of time quantum. If the quantum is very large, RR algorithm works the same as the FCFS algorithm. If the quantum is very small, RR algorithm

makes the user feels processor sharing between multiple processes very fast. Average waiting time is high because of FCFS policy and context switching [1].

D. Priority Based Scheduling

Priority scheduling executes processes based on their priority which may be assigned by the system or by the user himself [3]. Processes with the high priority are executed first and those with low priorities are executed next [6]. Processes with equal priority values are executed using FCFS approach [1].

E. Multilevel Queues (MLQ) Scheduling

It is a complex scheduling technique in which workload is divided among multiple queues employing different schedulers on different queues. Division of workload might be classified as system processes, interactive programs, batch jobs etc [2]. Each high priority queue contains foreground processes which have priority over lower priority queues which contain background processes. Processes keep moving between these queues depending on the scheduler employed on the particular queue [1].

IV. SJRR CPU SCHEDULING ALGORITHM

Shortest Job Round Robin (SJRR) is preemptive in nature. It sorts all incoming processes based on their burst time in ascending order in the ready queue. Next it uses the time quantum to execute processes. If time quantum expires before the process execution then CPU is preempted and given to the next shorter waiting process in the queue. The preempted process is then placed at the end of the ready queue. The average waiting time and average turnaround time obtained from SJRR is better than existing CPU scheduling algorithms.

SJRR is fair in scheduling and effective in time sharing environment. In SJRR scheduling, CPU is given to each process for equal time period, no process has to wait for long time for the CPU. The explicit working of the SJRR algorithm is discussed below:

- Take list of processes, their burst time, arrival time and time quantum.

- Arrange processes and their relative burst time in ascending order using any sorting technique.

- Iterate through the given list of processes to find the processes having maximum burst time and initialize waiting time of each process with zero.

- If number of processes are odd then
  - Take burst time of the middle process and assign this value to the time quantum

- Else
  - Take the average of burst time of two middle most processes and assign this value to the time quantum

- Find maximum number of time each process will execute by dividing maximum burst time with time quantum then add one in the result.

- Initialize an array with zero that is used for storing the burst time that has been completed.

- Iterate through the given list of processes.
  - Initialize a variable with zero that is used as a counter.
  - Iterate until the burst time of the process is greater than zero.
  - If burst time is greater than or equal to time quantum then
    - Store remaining burst time
    - Store completed burst time
    - Increment counter
  - Else
    - Store completed burst time
    - Assign zero to burst time variable
    - Increment counter
  - Assign value of counter minus one in counter array

- Iterate through the list of processes
  - Iterate through the length of counter array
    - If value of variable used for counter is equal to the counter of processes then
      - Iterate through the process coming from the list of processes
      - If value of that process is not equal to the upcoming process then
        - Add the waiting time of the upcoming process with the burst time completed.
    - Else
      - Iterate through the list of processes
      - If that process is not equal to the upcoming process then
        - Add waiting time of the upcoming process with the

burst time completed.

- Iterate through the list of processes
  - Add total waiting time with waiting time of each process to find total waiting time
  - Add burst time and waiting time of each process to find turnaround time
  - Add total turnaround time and turnaround time of each process to find total turnaround time
- Average waiting time is calculated by diving total waiting time with total number of processes.
- Average turnaround time is calculated by dividing total turnaround time with total number of processes.

V. PSEUDO CODE

```
burst ← 0
max ← 0
temp ← 0
total_tatime ← 0.0
tw_time ← 0.0
avg_wt ← 0.0
avg_tatime ←0.0

For i ← process-1 to 0
    For j ← 1 to process
        IF btime[j-1] > btime[j]
            temp ← btime[j-1]
            btime[j-1] ← btime[j]
            btime[j] ← temp
            ptemp ← proname[j-1]
            proname[j-1] ← proname [j]
            proname[j] ← ptemp

For i ← 0 to process
    btime[i] ← bu[i]
    proname[i] ← pname[i]
    IF max < btime[i]
        max ← btime[i]
    b[i] ← btime[i
    wtime[i] ← 0

IF process%2!=0
        mid← (process+1)/2
        t← btime[mid]
ELSE
        mid ← process/2
        t← (btime[mid]+btime[mid+1])/2

dim ← max/t + 1

For i ←0 to process
    For j ← 0 to dim
        r [i,j] ← 0
```

```
For i ← 0 to process
     j ←0
   While btime[i]>0
          IF btime[i] >=t
               btime[i] ←btime[i]-t
               r[i,j] ← t
               j ← j+1

          ELSE
               r[i,j] ← btime[i]
               btime[i] ← 0
               j←j+1
counter[i] ← j-1

For j ←0  to process
    For i ← 0 to counter [j]
        IF i = = counter[j]
          For k← 0 to j
             IF k != j
                 wtime[j] ← wtime[j] + r[k,i]

        ELSE
          For k ←o to process
             IF k != j
                 wtime[j] ← wtime[j] +r[k, i]

For j ←0 to process
    tw_time ← tw_time + wtime[j]
    tatime[j] ← b[j] + wtime[j]
    total_ tatime ← total_tatime+ tatime[j]

avg_wt ← tw_time / process
avg_tatime ← total_tatime/ process
```

VI.   SIMULATION DESIGN

The simulation provides an interactive GUI interface to the user through which a user can input data related to different processes then applying different algorithms based on the algorithm choices given in the simulator. The simulator employs .NET framework using C# on the front end and Microsoft Access for database at back end. Both front end and back end features are discussed below:

A.   Front End Features

The front end of the simulator has been built on .Net framework using C# as the coding language. It comprises of one parent screen and two child screens. The parent screen or the main window provides interactive GUI to the user which is used to choose two algorithms from the drop down list that are to be compared.

1)   First Screen Interface
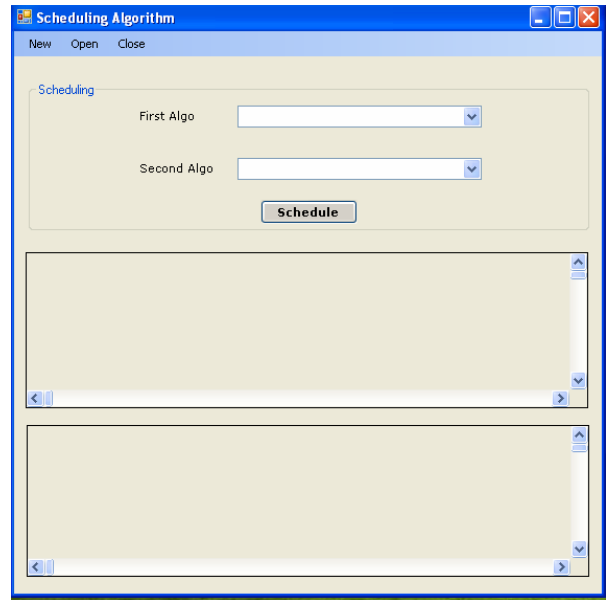
Snapshot 1: First Screen Interface

The Snapshot 1 of the parent screen illustrated above contains two drop down lists namely 'First Algo' and 'Second Algo', one for each algorithm respectively. Two algorithm choices are entered by the users that are to be compared. After choosing two algorithms the user then clicks the 'New' option. This selection is depicted below in Snapshots 2 & 3. These Snapshots illustrate how choices are entered and the selection choices are available. Same choice of algorithms in both dropdown lists is restricted.
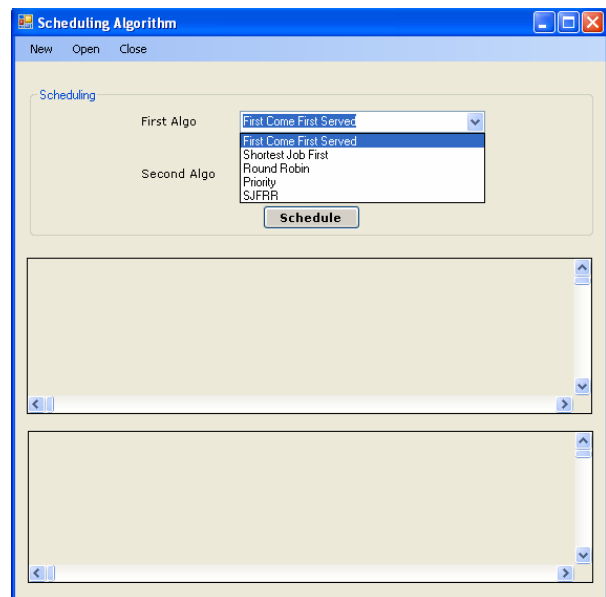
Snapshot 2: First Screen Interface with selection of First Algo
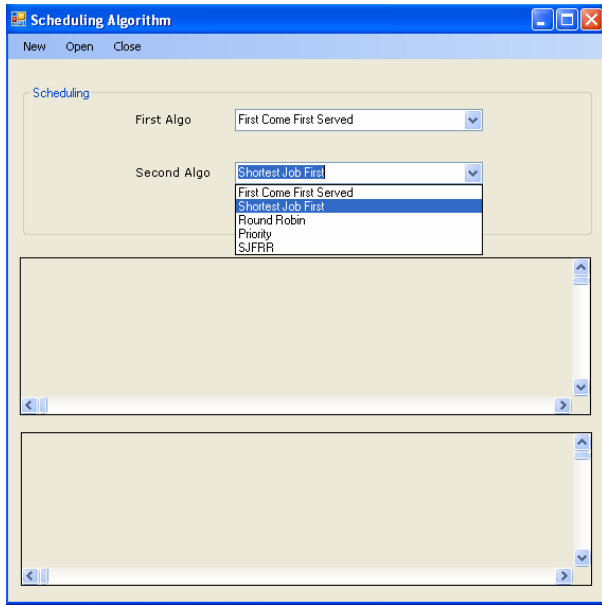
Snapshot 3: First Screen Interface with selection of Second Algo

Whenever in anyone of the given dropdown list, if Round Robin algorithm get selected another window pops up. This window is the Time Quantum Screen which will ask the user to enter the desired time quantum for the process set. All the processes will apply the same time quantum on all the processes for Round Robin algorithm.
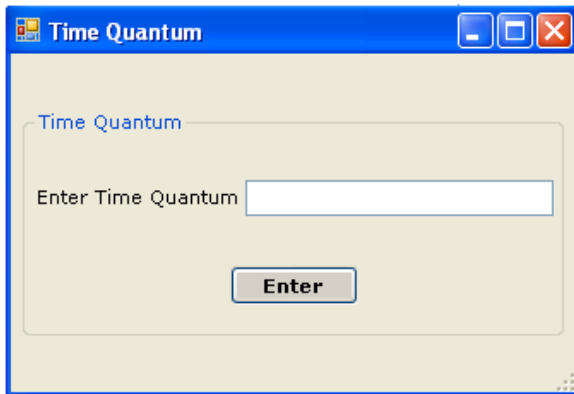
Snapshot 4: Time Quantum Screen

2) Second Screen Interface

This screen will open when 'New' option is clicked. It enables the users to enter the process set containing different processes each having its own specific process name, arrival time, burst time and priority. This set is used by both the chosen algorithms for applying their own techniques on the same process set. Given below is a snapshot which depicts this input in take.
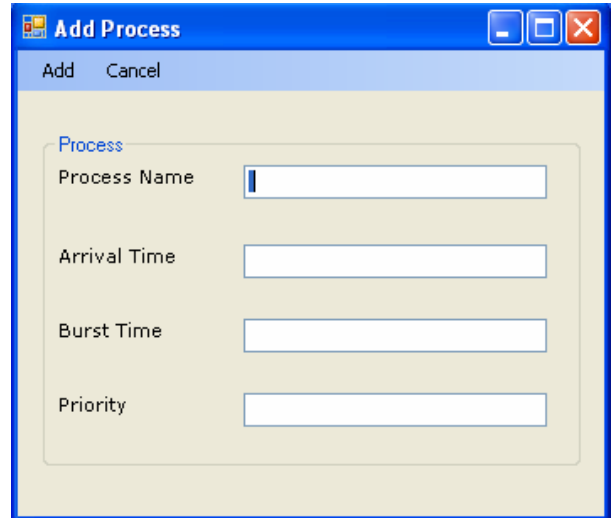
Snapshot 5: Second Screen Interface

After giving input for each process in the process set, then click the 'Add' option in the menu bar provided in the screen. Keep adding the processes and their specifics until all desired processes have been entered. After this close this screen which transfer the control to the first screen. Here, the user then clicks the 'Schedule' button which will apply the chosen algorithm techniques on the same process set separately. The first panel given at the bottom of the First Screen shows the Gantt chart results for the first algorithm that was entered while the second panel shows the Gantt chart results of the second algorithm.
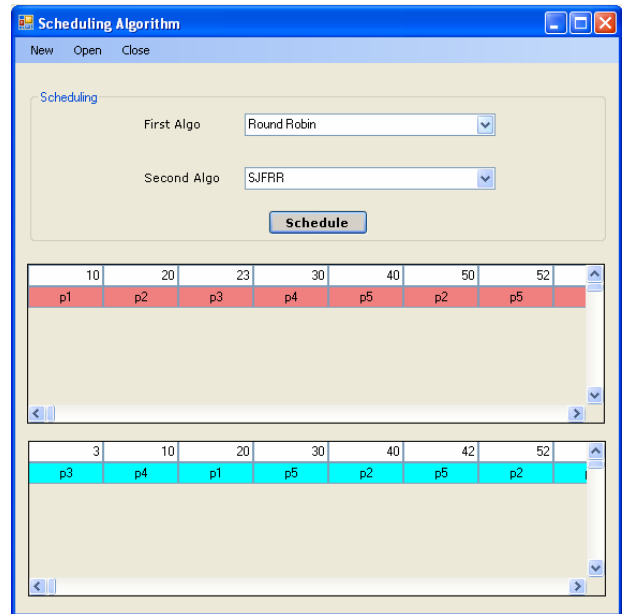
Snapshot 6: First Screen with Gantt Chart

B. Back End Features

Microsoft Access is employed at the back end which stores the processes, their arrival time, burst time and priorities entered by the user. The data field name and the data type used for storing and manipulating these values include:

- ID = Auto number
- Process = Text
- Arrival Time = Number
- Burst Time = Number
- Priority = Number

The ID number will be generated by the database at random as the processes get entered. The next data field is the Process name which can contain any string as named by the user. The Arrival Time takes numeric values for recording the arrival time of the processes. The Burst Time is the time which the process needs from the CPU for the completion of its task and has numeric data type. The priority of processes is also numeric which stores values and decide the priority of each process entered.

Snapshot 7: Database Screen

VII. COMPARISON OF SJRR WITH OTHER CPU SCHEDULING ALGORITHMS AND RESULTS

To compare the performance of SJRR, it was implemented with some other existing CPU scheduling algorithms. By taking lists of processes, processes are scheduled using different CPU scheduling algorithms. Average waiting time and average turnaround time was noted down for each. Results of average waiting time and average turnaround time for each algorithm are illustrated in the graphs below which depict behavior of each algorithm against others using a specific set of processes which random quantum size used per process set.

A. Comparison of SJRR with FCFS

Along the x-axis we have different set of processes that were chosen for SJRR and FCFS scheduling algorithms. Along y-axis we have the average waiting time and average turn around time spend by these set of processes while using these two scheduling approaches as given in Figure 1 and Figure 2 respectively. As depicted in the graph given in Figure 1, the difference between average waiting time (AWT) of FCFS and SJRR is quite large which makes SJRR an efficient algorithm with respect to the average waiting time. Same behavior was observed in the graph given in Figure 2, when average turnaround times (ATT) of the two algorithms were compared. The difference between the two plots, that is, plot of AWT and ATT; are quite negligible with small set of processes but the difference becomes noticeable when the set of processes grow large. Hence SJRR has better AWT and ATT readings as compared to FCFS.
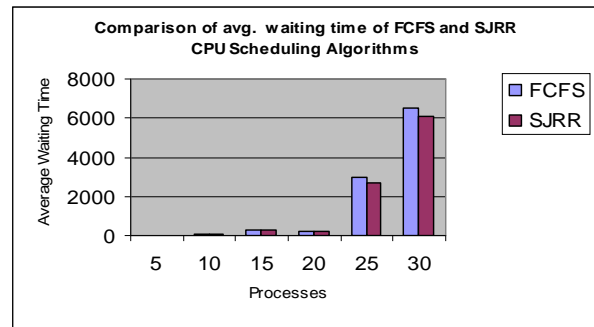
Figure 1: AWT of FCFS & SJRR Scheduling Algorithms
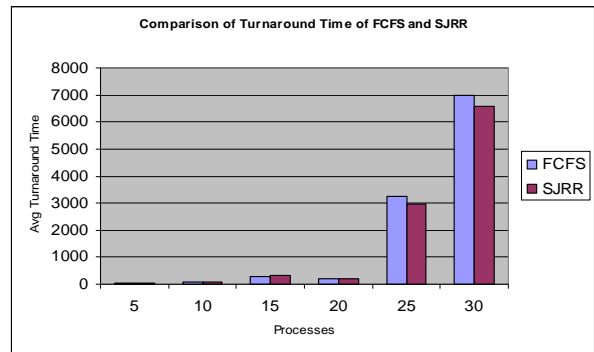
Figure 2: ATT of FCFS & SJRR Scheduling Algorithms

B. Comparison of SJRR with SJF

Along the x-axis we have different set of processes that were chosen for SJRR and SJF scheduling algorithms. Along y-axis we have the average waiting time and average turn around time spend by these set of processes while using these two scheduling approaches as given in Figure 3 and Figure 4 respectively. As depicted in the graphs given below, the AWT and ATT of SJRR was greater than AWT and ATT of SJF. SJRR showed negligible difference with SJF for small number of processes. But this difference starts becoming noticeable when set of processes grow above 20 processes.
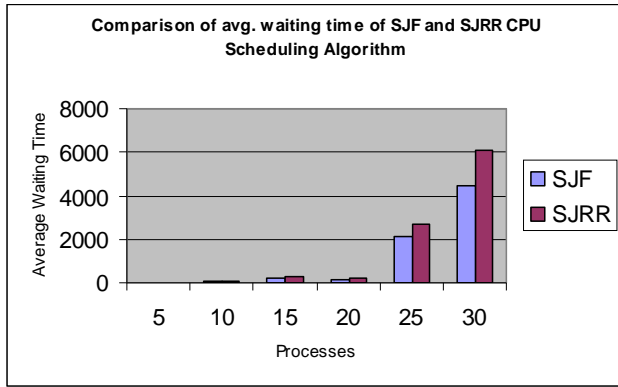
Figure 3: AWT of SJF & SJRR Scheduling Algorithms
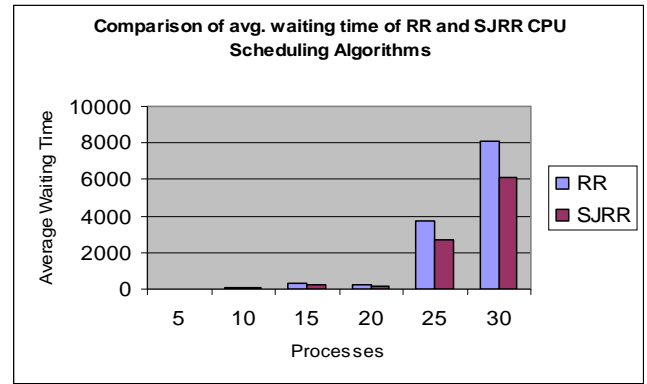
Figure 5: AWT of RR & SJRR Scheduling Algorithms
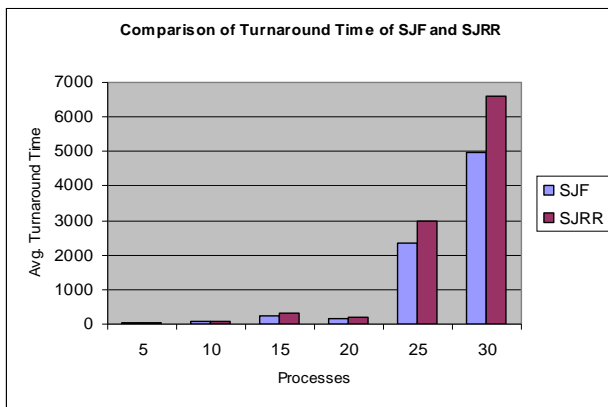
Figure 4: ATT of SJF & SJRR Scheduling Algorithms
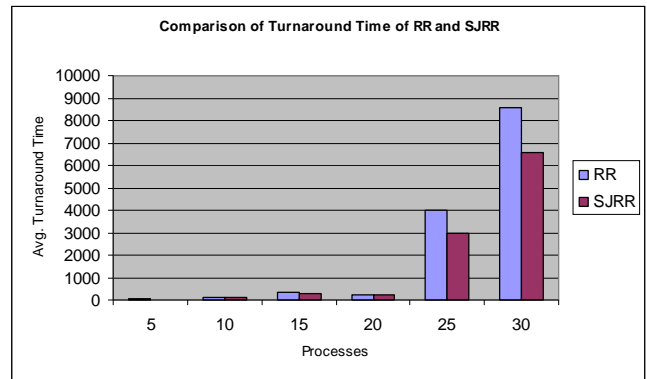
Figure 6: ATT of RR & SJRR Scheduling Algorithms

C. *Comparison of SJRR with RR*

Along the x-axis we have different set of processes that were chosen for SJRR and RR scheduling algorithms. Along y-axis we have the average waiting time and average turnaround time spend by these set of processes while using these two scheduling approaches as given in Figure 5 and Figure 6 respectively. SJRR shows negligible difference with AWT and ATT of RR when the set of processes is small. RR gave better AWT and ATT readings when time quantum was small for a set of processes. But SJRR produced much better AWT and ATT results as compared to RR when the time quantum was increased. That is, for large quantum size, SJRR gives much better results as compared to RR.

D. *Comparison of SJRR with Priority*

Along the x-axis we have different set of processes that were chosen for SJRR and Priority scheduling algorithms. Along y-axis we have the average waiting time and average turn around time spend by these set of processes while using these two scheduling approaches as given in Figure 7 and Figure 8 respectively. As shown in the graphs SJRR shows better readings of AWT and ATT as compared to Priority. There is a small difference between the results of AWT and ATT of SJRR and the results of AWT and ATT of Priority.
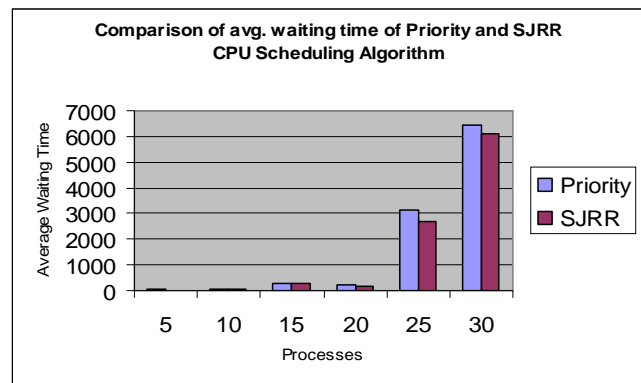
Figure 7: AWT of Priority & SJRR Scheduling Algorithms
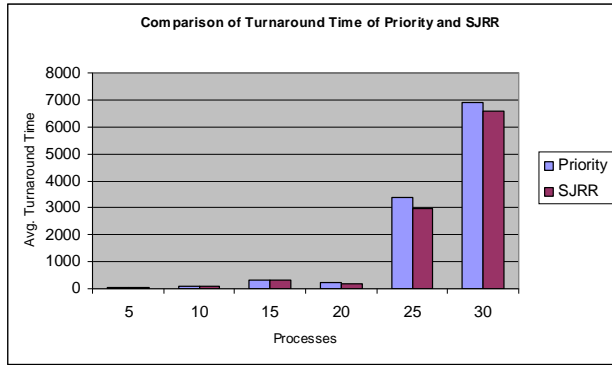
Figure 8: ATT of Priority & SJRR Scheduling Algorithms

E. Comparative Analysis of Above CPU Scheduling Algorithms

Waiting time and Turn Around time are some of the factors that can be used to check efficiency of a CPU Scheduling algorithm. Waiting time gives the amount of time a process has spent in the ready queue waiting CPU execution. The more the time spent by a process in the ready queue the more time will be taken by the CPU to execute processes queued behind that process which will retard CPU efficiency. Similarly turn around time is the amount of time taken by a CPU to execute a process. The more this time is taken the slower the CPU gets making other processes starve infinitely.

To check the performance of proposed algorithm, i.e. SJRR, we took six process sets, each with different characteristics and made comparisons of these. First process set contains 5 processes, second set contains 10 processes, third set contains 15 processes, fourth set contains 20 processes, fifth set contains 25 and sixth set contains 30 processes. Each process has its own specific CPU burst time, arrival time, priority number and time quantum.

Performance metrics for the CPU scheduling algorithms are based on two main factors – Average Waiting Time and Average Turnaround Time. Table 1 shows the Average Waiting Time for each process set and Table 2 shows the Average Turnaround Time for each process set.

Processes	FCFS	SJF	RR	Priority	SJRR
5	28	13	23	32.8	15
10	69.1	51.2	86.2	84	68.2
15	260.47	213.67	323	266.47	266.8
20	177.15	153.1	224.45	205.5	185.5
25	2992.8	2096	3751.2	3126	2712.4
30	6534	4495	8116	6428	6122.5

Table 1: Experimental Results for Average Waiting Time

Processes	FCFS	SJF	RR	Priority	SJRR
5	40.2	25.2	35.2	45	27.2
10	87.4	69.5	104.5	102.3	86.5
15	299.5	252.73	362.07	305.5	305.9
20	196.85	172.8	244.15	225.2	205.2
25	3254	2358	4013.2	3388	2974.4
30	6999	4960	8581	6893	6587.5

Table 2: Experimental Results for Average Turnaround Time

Below is the graphical depiction of Table 1 and Table 2. Figure 9 illustrates cumulative comparison of AWT of SJRR with AWT of FCFS, SJF, RR and Priority scheduling algorithms. Figure 10 illustrates cumulative comparison of ATT of SJRR with FCFS, SJF, RR and Priority scheduling algorithms. Analysis of both plots show that AWT and ATT of SJRR is better than AWT and ATT of FCFS, RR and Priority but SJRR shows an increase in AWT and ATT as compared to that of SJF. Using these analysis results it is concluded that SJRR show a marked improvement in AWT and ATT perspectives as compared to those of FCFS, RR and Priority scheduling algorithms.
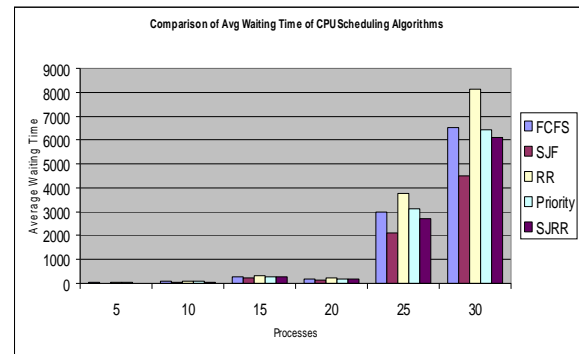
Figure 9: Average Waiting Time of CPU Scheduling Algorithms
(Comparative Analysis)
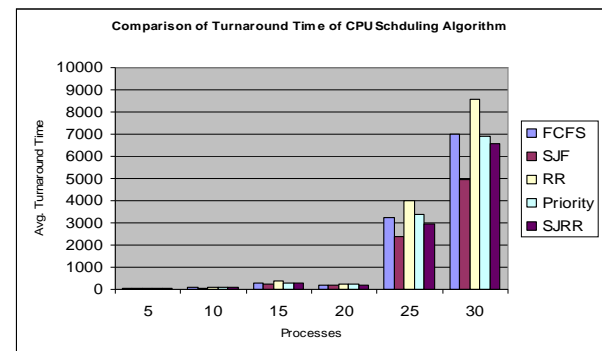
Figure 10: Average Turnaround Time of CPU Scheduling Algorithms
(Comparative Analysis)

VIII. CONCLUSION

The paper presents a new CPU scheduling algorithm called SJRR CPU Scheduling Algorithm. Also presented in this paper are the simulation interface and its working which interactively takes input from the user and compares the process set against different algorithm pairs. The results of the simulation for different process sets using different scheduling algorithms has been presented graphically in this piece of work. The last half of the paper provides analytical result with each set of graph. From the above graphs and results, it is clear that SJRR is more efficient than First Come First Served, Round Robin and Priority although it is less efficient than Shortest Job First. In future, an enhanced and more efficient version of this algorithm along with enhanced simulation will be presented.

REFERENCES

[1] Abraham Silberschatz , Peter Baer Galvin, Greg Gagne, "Operating System Concepts",Sixth Edition.

[2] Milan Milenkovic, "Operating System Concepts and Design", McGRAW-HILL, Computer Science Series, Second Edition.

[3] H.M.Deitel, "Operating Systems", Pearson Education, Second Edition.

[4]   http://en.wikipedia.org/wiki/Scheduling (computing).

[5] Sukanya Suranauwarat, "A CPU Scheduling Algorithm Simulator", 37th ASEE/IEEE Frontiers in Education Conference.

[6] Akhtar Hussain, "Optimized Performance of CPU Scheduling", Master's Thesis, National University of Science and Technology.

[7] M Gary Nutt, "Operating Systems- A Modern Perspective", Second Edition, Pearson Education, 2000

Robust Resilient Two Server Password Authentication Vs Single Server

T.S.THANGAVEL
Department of M.Sc (IT)
K.S.Rangasamy College of Technology
Tiruchengode
Tamilnadu
India

Dr.A.KRISHNAN
Department of Electronic and Communication Engg.
K.S.Rangasamy College of Technology
Tiruchengode
Tamilnadu
India

Abstract

The authentication system stores the password in a Central Server, and the possibility for the intruder to obtain the password is very easy and can gain access to the contents of the user. For the purpose of authentication, the multi-server systems we proposed to communicate with one or all of the servers. It requires high communication bandwidth at the same time is not easy to maintain and also the protocols are highly expensive. The Two Server Authentication System avoids this problem, which uses the passwords and the session keys, rather than performing the cryptographic techniques. It consists of two servers, the front end and the back end server. The front end server communicates with the user, whereas the back end control server is only visible to the service server. These two servers are responsible for the authentication. The password is split into two words, which is one with the service server and the other with the control server. Both the servers are validated during the form validation process. The system is suitable for both the computation and communication system. The servers are also used for the multiple clients and also for the single server systems.

Keywords: Password-Authentication, Two Servers password, Cryptosystem, single sever Secure Password, Service sever, control server.

I. INTRODUCTION

The multi-user systems require the users to provide their passwords along with their user identification. The password serves to authenticate the ID of the individual logging on to the system. This is required to determine if the user is authorized to gain access to the system. This ID also determines the privileges accorded to the user. The short secrets are convenient, particularly for an increasingly mobile user population. Many users are interested in employing a variety of computing devices with different forms of connectivity and different software platforms. Such users often find it convenient to authenticate by means of passwords and short secrets, to recover lost passwords by answering questions, and to make similar use of relatively weak secrets.

Most password-based user authentication systems place total trust on the authentication server where passwords or easily derived password verification data are stored in a central database. These systems could be easily compromised by offline dictionary attacks initiated at the server side. Compromise of the authentication server by either outsiders or insiders subjects all user passwords to exposure and may have serious problems. To overcome these problems in the single server system many of the systems has been proposed such as multi-server systems, public key cryptography and password systems, threshold password authentication systems, two server password authentication systems.

The proposed work continues the line of research on the two-server paradigm in [10], [11], extend the model by imposing different levels of trust upon the two servers, and adopt a very different method at the technical level in the protocol design. As a result, we propose a practical two-server password authentication and key exchange system that is secure against offline dictionary attacks by servers when they are controlled by adversaries. The proposed scheme is a password-only system in the sense that it requires no public key cryptosystem and, thus, no PKI. This makes the system very attractive considering PKIs are proven notoriously expensive to deploy in real world. Moreover, the proposed system is particularly suitable for resource constrained users due to its efficiency in terms of both computation and communication. The paper work, generalize

the basic two-server model to architecture of a single back-end server supporting multiple front-end servers and envision interesting applications in federated enterprises.

II.   LITERATURE REVIEW

Public key techniques are absolutely necessary to make password systems secure against offline dictionary attacks, whereas the involvement of public key cryptosystems under a PKI (e.g., public key encryption and digital signature schemes) is not essential. There are two separate approaches to the development of secure password systems one is a combined use of a password and public key cryptosystem under a PKI, and the other is a password only approach. In these systems, the use of public keys entails the deployment and maintenance of a PKI for public key certification and adds to users the burden of checking key validity. To eliminate this drawback, password-only protocols (password authenticated key exchange or PAKE) have been extensively studied, e.g., [2], [3], [4]. The PAKE protocols do not involve any public key cryptosystem under a PKI and, therefore, are much more attractive for real-world applications. Any use of public key cryptosystem under a PKI in a password authentication system should be avoided since, otherwise, the benefits brought by the use of password would be counteracted to a great extent.

Most of the existing password systems were designed over a single server, where each user shares a password or some password verification data (PVD) with a single authentication server (e.g., [2], [3], [4]). These systems are essentially intended to defeat offline dictionary attacks by outside attackers and assume that the sever is completely trusted in protecting the user password database. Unfortunately, attackers in practice take on a variety of forms, such as hackers, viruses, worms, accidents, mis-configurations, and disgruntled system administrators. As a result, no security measures and precautions can guarantee that a system will never be penetrated. Once an authentication server is compromised, all the user passwords or PVD fall in the hands of the attackers, who are definitely effective in offline dictionary attacks against the user passwords. To eliminate this single point of vulnerability inherent in the single-server systems, password systems based on multiple servers were proposed. The principle is distributing the password database as well as the authentication function to multiple servers so that an attacker is forced to compromise several servers to be successful in offline dictionary attacks.

The system in [6], believed to be the first multiserver password system, splits a password among multiple servers. However, the servers in [6] need to use public keys. An improved version of [6] was proposed in [7], which eliminates the use of public keys by the servers. Further and more rigorous extensions were due to [8], where the former built a t-out-of-n threshold PAKE protocol and provided a formal security proof under the random oracle model [5] and the latter presented two provably secure threshold PAKE protocols under the standard model. While the protocols are theoretically significant, they have low efficiency and high operational overhead. In these multi-server password systems, either the servers are equally exposed to the users and a user has to communicate in parallel with several or all servers for authentication, or a gateway is introduced between the users and the servers.

Recently, Brainard et al. [1] proposed a two-server password system in which one server exposes itself to users and the other is hidden from the public. While this two-server setting is interesting, it is not a password-only system: Both servers need to have public keys to protect the communication channels from users to servers. As we have stressed earlier, this makes it difficult to fully enjoy the benefits of a password system. In addition, the system in [1] only performs unilateral authentication and relies on the Secure Socket Layer (SSL) to establish a session key between a user and the front-end server. Subsequently, Yang et al. [9] extended and tailored this two-server system to the context of federated enterprises, where the back-end server is managed by an enterprise headquarters and each affiliating organization operates a front-end server. An improvement made in [9] is that only the back-end server holds a public key. Nevertheless, the system in [9] is still not a password-only system.

III. MODES OF SERVER PASSWORD AUTHENTICATION MODELS

In the single-server model as shown in fig1, where a single server is involved and it keeps a database of user passwords. Most of the existing password systems follow this single-server model, but the single server results in a single point of vulnerability in terms of offline dictionary attacks against the user password database.
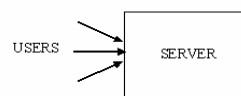

Fig 1: Single Server Password model

In the multi-server model, the server side comprises multiple servers for the purpose of removing the single point of vulnerability, the servers are equally exposed to users and a user has to communicate in parallel with several or all servers for authentication. The main problem with the plain multi-server model is the demand on communication bandwidth and the need for synchronization at the user side since a user has to engage in simultaneous communications with multiple servers. This may cause problems to resource-constrained mobile devices such as hand phones and PDAs.
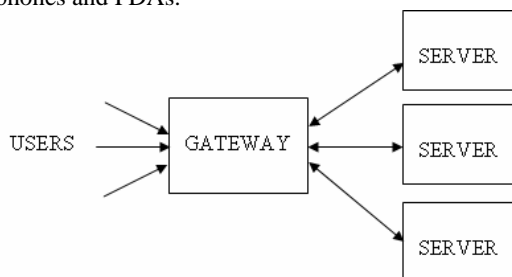
Fig 2: Gateway Augmented Multi-server model

In the gateway augmented multi-server model as shown fig2, gateway is positioned as a relaying point between users and servers and a user only needs to contact the gateway. Apparently, the introduction of the gateway removes the demand of simultaneous communications by a user with multiple servers as in the plain multi-server model. However, the gateway introduces an additional layer in the architecture, which appears "redundant" since the purpose of the gateway is simply to relay messages between users and servers, and it does not in any way involve in service provision, authentication, and other security enforcements. From security perspective, more components generally imply more points of vulnerabilities.
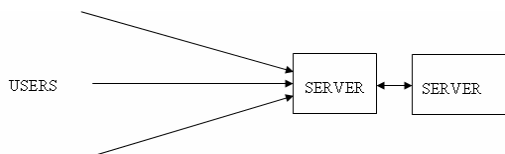
Fig 3: Two server model

The two-server model comprises two servers at the server side, one of which is a public server exposing itself to users and the other of which is a back-end server staying behind the scene; users contact only the public server, but the two servers work together to authenticate users. The differences between the two-server model and the earlier multi-server models are

a) In the two-server model, a user ends up establishing a session key only with the public server, and the role of the back-end server is merely to assist the public server in user authentication, while in the multi-server models, a user establishes a session key (either different or the same) with each of the servers.

b) From a security point of view, servers in the multi-server models are equally exposed to outside attackers (recall that the gateway in the gateway augmented multi-server model does not enforce security), while in the two-server model, only the public server faces such a problem. This improves the server side security and the overall system security in the two-server model.

In two server model, different levels of trust upon the two servers with respect to outside attackers can be made. The back-end server is more trustworthy than the public server. This is logical since the back-end server is located in the back-end and is hidden from the public, and it is thus less likely to be attacked. Two-server model has successfully eliminated drawbacks in the plain multi-server model (i.e., simultaneous communications between a user and multiple servers) and the gateway augmented multi-server model (i.e., redundancy) while allowing us to distribute user passwords and the authentication functionality to two servers in order to eliminate a single point of vulnerability in the single-server model. As a result, the two-server model appears to be a sound model for practical applications.

The existing systems upon the two-server model are not suffice, in turn motivated to present a password-only system over the two-server model. In the proposed system, the public server acts as a service server that provides application services, while the back-end server is a control server whose sole purpose is to assist the service server in user authentication (the service server, of course, also participates in user authentication). In the plain multi-server model and the gateway augmented multi-server model, several or all servers equally participate in service provision as well as user authentication, which is implied by the fact that a user negotiates a session key with each server. The two-server model is generalized to architecture that a control server supports multiple service servers.

IV. FUNCTIONAL ARCHITECTURE OF TWO SERVER PASSWORD AUTHENTICATION SYSTEM

Three types of entities are involved in our system, i.e., users, a service server (SS) that is the public server in the two server model, and a control server (CS) that is the back-end server. In this setting, users only communicate with SS and do not

necessarily know CS. For the purpose of user authentication, a user U has a password which is transformed into two long secrets, which are held by SS and CS, respectively. Based on their respective shares, SS and CS together validate users during user login. CS is controlled by a passive adversary and SS is controlled by an active adversary in terms of offline dictionary attacks to user passwords, but they do not collude (otherwise, it equates the single-server model).

A passive adversary follows honest-but-curious behavior, that is, it honestly executes the protocol according to the protocol specification and does not modify data, but it eavesdrops on communication channels, collects protocol transcripts and tries to derive user passwords from the transcripts, moreover, when an passive adversary controls a server, it knows all internal states of knowledge known to the server, including its private key (if any) and the shares of user passwords. In contrast, an active adversary can act arbitrarily in order to uncover user passwords. Besides, we assume a secret communication channel between SS and CS for this basic protocol. This security model exploits the different levels of trust upon the two servers. This holds with respect to outside attackers. As far as inside attackers are concerned, justifications come from our application and generalization of the system to the architecture of a single control server supporting multiple service servers, where the control server affords and deserves enforcing more stringent security measurements against inside attackers. The back-end server is strictly passive and is not allowed to eavesdrop on communication channels, while CS in our setting is allowed for eavesdropping.
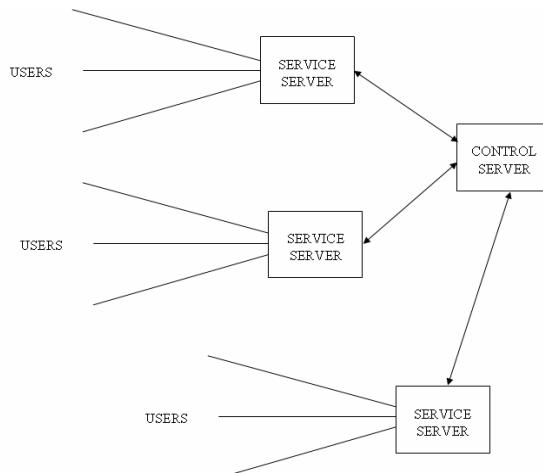
The user contacts only the service server but both the control and service servers are responsible for the authentication of the user. The user has a password which is transformed into two long secrets which are held by service server and control server. Both the system using their respective shares validate user during the login. The servers compute function to verify the user and finally a session key is being established between the user and service server for the confirmation of the user and the server. The service server (Fig 5) which is an active adversary acts arbitrarily to uncover the passwords and could control the corruption of the password.

The user can use the same password to register to different servers, the service server connect either to distinct control servers or to the same control server. It makes the system user friendly. The system could be adapted to any existing FTP and web applications that are available today by adding a control server.

In our experimental implementation, a password is split into two random numbers. Therefore, a user can use the same password to register to different service servers; they connect either to distinct control servers or to the same control server.

This is a highly desirable feature since it makes the system user friendly. A big inconvenience in the traditional password systems is that a user has to memorize different passwords for different applications. The system has no compatibility problem with the single-server model. This is of importance, as most of the existing password systems use a single server.

Fig 4: Generalized Two Server Architecture of a single control server with multiple service server

V EXPERIMENTAL PERFORMANCE EVALUATION
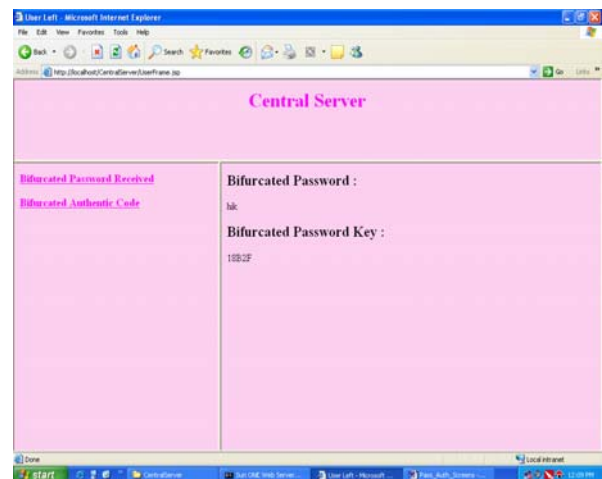

Fig 5: Service Server Key generation for user password (bifurcated)

The generalization as well as the applications of the two-server password system well support the underlying security model, in the sense that the enterprise headquarter naturally assume adequate funds and strong security expertise and, therefore, affords and is capable of maintaining a highly trustworthy control server against both inside attackers and outside attackers. Without the concern of a single point of vulnerability, affiliating organizations that operate service servers are offloaded to some extent from strict security management, so they can dedicate their limited expertise and resources to their core competencies and to enhancing service provision to the users. From the perspective of users, they are able to assume the higher creditability of the enterprise while engaging in business with individual affiliating organizations.

A. Performance Measure

The exponentiations dominate each party's computation overhead, the two server password authentication system only count the number of exponentiations as the computation performance. The digits before "/" denote the total number of exponentiations performed by each party, and the digits following "/" denote the number of exponentiations that can be computed offline. One round is a one-way transmission of messages. The proposed two protocols demonstrate performance quite efficient in terms of both computation and communication to all parties. Take U, for example, it needs to calculate 3 and 4 exponentiations in the two protocols, respectively, and 2 of them can be performed offline. This means U only computes 1 and 2 exponentiations in real time in the respective protocols, the communication overhead for U is particularly low in terms of both bits and rounds. The table 1 listed below indicates the computation performance in terms of time and success rate (number of rounds) of the two server password authentication and single server authentication

Table 1: Performance measure on Two server and Single server password authentication scheme

Scheme	Time of Authenticity	Success rate
	(milliseconds)	%
Two server password authentication	10	96
Single server	8	87

B. Implementation

The implementation procedure is discussed below:

- Password is split into two random numbers.
- User use the same password to register to different service servers
- They connect either to distinct control servers or to the same control server.
- The service server, an active adversary, acts arbitrarily to uncover the passwords and control the corruption of the password,
- Control server, a passive adversary, acts according to the authentic function.
- To initiate a request for service, User U sends his identity together with a service request to SS in M1.
- SS first relays the request to CS by sending the user ID in M2.,
- Then selects a random number b1 and computes B1 using his password share 1.
- Upon receiving M2, CS chooses a random number b2 and computesB2 using his password share 2.
- CS then sends B2 in M3 to SS.
- Upon reception of B2, SS computes and sends B to U in M4.
- After receiving M4, U selects and computes A and Su.
- U then sends A and Su to SS in M5.
- Getting the message, SS computes S1 and sends S1, A and Su to CS in M6.
- Upon receipt of M6, CS computes S2 and checks whether Su.
- If it holds, CS is assured of the authenticity of U, and continues the protocol by sending S2 to SS in M7 otherwise, CS aborts the protocol.
- Assuming SS receives S2 in M7, it checks whether Su.
- If it holds, SS is convinced of the authenticity of U.
- At this stage, both servers have authenticated U.
- SS then computes and sends Ss to U in M8 and afterward computes a session key K otherwise, SS aborts the protocol.
- Upon receiving M8, U checks if h(0, Sou) = Ss.
- If it holds, U has validated the servers and then computes a session key K otherwise, U aborts the protocol.

C.Discussions

With two-server password system, single point of vulnerability, is totally eliminated. Without compromising both servers, no attacker can find user passwords through offline dictionary attacks. The control server being isolated from the public, the chance for it being attacked is substantially

minimized, thereby increasing the security of the overall system. The system is also resilient to offline dictionary attacks by outside attackers. This allows users to use easy to remember passwords and still have strong authentication and key exchange. The system has no compatibility problem with the single-server model. The generalization of the two-server password system well supports the underlying security model. In reality, adversaries take on a variety of forms and no security measures and precautions can guarantee that a system will never be penetrated. By avoiding a single point of vulnerability, it gives a system more time to react to attacks. The password-based authentication and key exchange system that is built upon a novel two-server model, where only one server communicates to users while the other server stays transparent to the public. Compared with previous solutions, our system possesses many advantages, such as the elimination of a single point of vulnerability, avoidance of PKI, and high efficiency.

VI. CONCLUSION

The Two-Server password authentication architecture consists of two servers, namely control server and service server. The control server is controlled by a passive adversary whereas the service server is controlled by an active adversary. The factor, vulnerability is eliminated in this process. Both servers are required without which the attacker cannot find the user passwords. The control server is isolated from the public. So the possibility of being it attacked is minimized hence the overall system is protected. The password is split into two random numbers. The user can use the same password for both the servers. Hence the overall system is user friendly. Both the inside attackers and the outside attackers cannot easily enter into the system. The two server system is highly used for practical applications.

In contrast to existing multi-server password systems, the two server system has great potential for practical applications. It can be directly applied to fortify existing standard single-server password applications, e.g., FTP and Web applications.

References

[1] J. Brainard, A. Juels, B. Kaliski, and M. Szydlo, "A New Two Server Approach for Authentication with Short Secrets," Proc. USENIX Security Symp., 2003.

[2] S. Bellovin and M. Merritt, "Encrypted Key Exchange: Password Based Protocols Secure against Dictionary Attacks," Proc. IEEE Symp. Research in Security and Privacy, pp. 72-84, 1992.

[3] S. Bellovin and M. Merritt, "Augmented Encrypted Key Exchange: A Password-Based Protocol Secure against Dictionary Attacks and Password File Compromise," Proc. ACM Conf. Computer and Comm. Security, pp. 244-250, 1993.

[4] M. Bellare, D. Pointcheval, and P. Rogaway, "Authenticated Key Exchange Secure Against Dictionary Attacks," Advances in Cryptology (Eurocrypt '00), pp. 139-155, 2000.

[5] M. Bellare and P. Rogaway, "Random Oracles are Practical: A Paradigm for Designing Efficient Protocols," Proc. ACM Computer and Comm. Security, pp. 62-73, 1993.

[6] W. Ford and B.S. Kaliski Jr., "Server-Assisted Generation of a Strong Secret from a Password," Proc. IEEE Ninth Int'l Workshop Enabling Technologies, 2000.

[7] D.P. Jablon, "Password Authentication Using Multiple Servers," RSA Security Conf., pp. 344-360, 2001.

[8] P. Mackenzie, T. Shrimpton, and M. Jakobsson, "Threshold Password-Authenticated Key Exchange," Proc. Advances in Cryptology (Eurocrypt '02), pp. 385-400, 2002.

[9] Y.J. Yang, F. Bao, and R.H. Deng, "A New Architecture for Authentication and Key Exchange Using Password for Federated Enterprises," Proc. 20th Int'l Federation for Information Processing Int'l Information Security Conf. (SEC '05), 2005.

[10] Yanjiang Yang, Robert H. Deng, and Feng Bao, "A Practical Password-Based Two Server Authentication and Key Exchange System," IEEE Transaction on Secure and Dependable Computing,Vol.3, No.2,April-June 2006

[11] C. Ellison, C. Hall, R. Milbert, and B. Schneier. Protecting secret keys with personal entropy. Journal of Future Generation Computer Systems, 16(4):311-318, February 2000.

[12] N. Frykholm and A. Juels. Error-tolerant password recovery. In P. Samarati, editor, 8th ACM Conference on Computer and Communications Security, pages 1-9. ACM Press, 2001.

Author's Profile

[1]**T.S.Thangavel** received the Bsc degree in Computer Science (Bharathiyar University) in 1991 and the Msc degree in computer science(Bharathidasan University) in 1993 and the Mphil degree in Computer Science (Bharathidasan university) in 2003. He is pursuing the PhD degree in department of science and humanities (Anna university). He is working as an assistant professor in MCA department at K.S.Rangasamy College of Technology, Tiruchengode

[2]**Dr. A. Krishnan** received his Ph.D degree in Electrical Engineering from IIT, Kanpur. He is now working as an Academic Dean at K.S.Rangasamy College of Technology, Tiruchengode and research guide at Anna University Chennai. His research interest includes Control system, Digital Filters, Power Electronics, Digital Signal processing, Communication Networks. He has been published more than 165 technical papers at various National/ International Conference and journals.

Effective MSE optimization in fractal image compression

A.Muruganandham
Sona College of Technology,
salem-05.,India.
muruga_salem@rediffmail.com,

Dr.R.S.D.wahida banu
Govt Engineering College
Salem-11, India
rsdwb@yahoo.com

Abstract- The Fractal image compression encodes image at low bitrate with acceptable image quality, but time taken for encoding is large. In this paper we proposed a fast fractal encoding using particle swarm optimization (PSO). Here optimization technique is used to optimize MSE between range block and domain block. PSO technique speedup the fractal encoder and preserve the image quality.

Keywords- mean square error (MSE) ,particle swarm optimization (PSO), fractal image compression (FIC), Iteration Function System (IFS)

I. INTRODUCTION

The idea of the fractal image compression (FIC) is based on the assumption that the image redundancies can be efficiently exploited by means of block self-affine transformations. The fractal transform for image compression was introduced in 1985 by Barnsley and Demko [1,2]. The first practical fractal image compression scheme was introduced in 1992 by Jacquin [3,4]. One of the main disadvantages using exhaustive search strategy is the very high encoding time. Therefore, decreasing the encoding time is an interesting research topic for FIC [3, 4].

An approach for decreasing encoding time is using the stochastic optimization methods such as genetic algorithm (GA). Some recent GA-based methods are proposed to improve the efficiency [5, 6]. The idea of special correlation of an image is used in these methods, which is of great interesting. While the chromosomes in GA consist of all range blocks which leads to high encoding speed and particular properties of natural images have never been used that will results in lose of visual effect in the retrieved image.

Other researchers focused on improvements of the search process to make it faster by tree structure search methods [12,13], parallel search methods [14,15] or quad tree partitioning of range blocks [9,16,].

In this paper, we present a fast fractal encoding using particle swarm optimization. The outline of the remaining part of this paper is as follows: Section II includes fractal image coding. Section III involves implementation of PSO. Section IV concerns the proposed fast fractal encoder using PSO, and in Section V experimental results are included. In Section VI, we present our conclusions.

II. FRACTAL IMAGE COMPRESSION ALGORITHM

The Iteration Function System (IFS) is the fundamental idea of fractal image compression in which the governing theorems are the Collage Theorem and the Contractive Mapping Fixed-Point Theorem [7]. The encoding unit of FIC for given gray level image of size N x N is $(N/L)^2$ of non-overlapping range blocks of size L x L which forms the range pool *R*. For each range block *v* in *R*, one search in the $(N - 2L + 1)^2$ overlapping domain blocks of size 2L x 2L which forms the domain pool *D* to find the best match. The parameters describing this fractal affine transformation of domain block into range block form the fractal compression code of *v*.

The parameters of fractal affine transformation *Φ* of domain block into range block is domain block coordinates-(*tx, ty)*, Dihedral transformation-*d*, contrast scaling-*p*, brightness offset-*q*. Flowchart for this fractal affine transformation is illustrated in fig. 1 and it is given by

$$\phi \begin{bmatrix} x \\ y \\ u(x,y) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & p \end{bmatrix} \begin{bmatrix} x \\ y \\ u(x,y) \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ q \end{bmatrix},$$

(1)

where the 2 x 2 sub-matrix $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ is one of the Dihedral transformations in (2)

$$T_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad T_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad T_2 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$T_3 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, T_4 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad T_5 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

$$T_6 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad T_7 = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

(2)
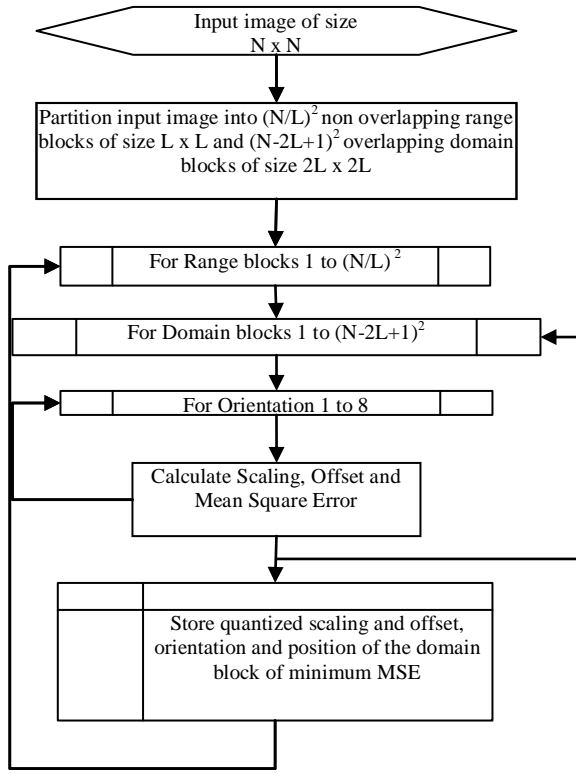
Fig. 1 fractal affine transformation of domain block into range block

The above parameters are found using the following procedure

1. the domain block is first down-sampled to L x L and denoted by u

2. The down-sampled block is transformed subject to the eight transformations Tk:k = 0,. . . ,7 in the Dihedral on the pixel positions and are denoted by u_k, k = 0,1, . . . ,7, where $u_0 = u$. The transformations T_1 and T_2 correspond to the flips of u along the horizontal and vertical lines, respectively. T_3 is the flip along both the horizontal and vertical lines. T_4, T_5, T_6, and T_7 are the transformations of T_0, T_1, T_2, and T_3 performed by an additional flip along the main diagonal line, respectively.

3. For each domain block, there are eight separate MSE computations required to find the index d such that

$$d = \arg\min\{ \text{MSE}((p_k u_k + q_k), v) : k = 0,1,\ldots,7\}$$
(3)

where

$$\text{MSE}(u,v) = \frac{1}{L^2} \sum_{i,j=0}^{L-1} (u(i, j = 0) - v(i, j))^2$$
(4)

Here, p_k and q_k can be computed directly as

$$p_k = \frac{L^2 \langle u_k, v \rangle - \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} u_k(i,j)\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} v(i,j)}{\left[L^2 \langle u_k, u_k \rangle - \left(\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} u_k(i,j)\right)^2 \right]};$$
(5)

$$q_k = \frac{1}{L^2}\left[\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} v(i,j) - p_k \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} u_k(i,j) \right].$$
(6)

4. As u runs over all of the domain blocks in D to find the best match, the terms t_x and t_y can be obtained together with d and the specific p and q corresponding this d, the affine transformation (1) is found for the given range block v.

In practice, t_x, t_y, d, p, and q can be encoded using $\log_2(N)$, $\log_2(N)$, 3, 5, and 7 bits, respectively, which are regarded as the compression code of v. Finally, the encoding process is completed as v runs over all of the $(N/L)^2$ range blocks in R.

Fig. 2 show the MSE vs. quantization parameter for an randomly selected range block of size 8 x 8 from 256 x 256 Lena image. From fig. 2, choosing 5 bits and 7 bits as quantization parameter for scale and offset value is justified.
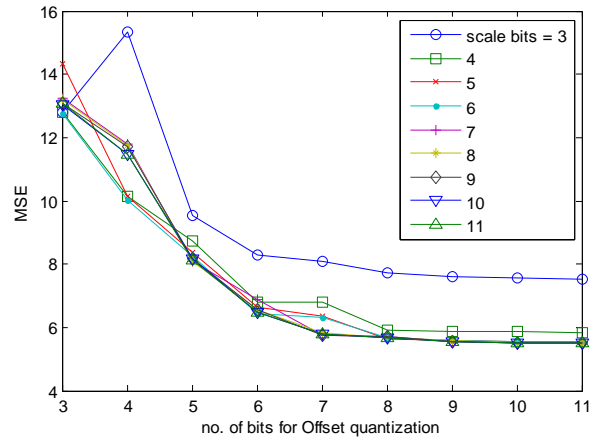
Fig. 2 MSE Vs. quantization parameter

To decode, chooses any image as the initial one and makes up the $(N/L)^2$ affine transformations from the compression codes to obtain a new image, and proceeds recursively. According to Partitioned Iteration Function Theorem (PIFS), the sequence of images will converge. According to user's application the stopping criterion of the recursion is designed. The final image is the retrieved image of fractal coding.

III. PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a population-based algorithm for searching global optimum. It ties to artificial life, like fish schooling or bird flocking, and has some common features of evolutionary

computation such as fitness evaluation. The original idea of PSO is to simulate a simplified social behavior [8, 9]. Similar to the crossover operation of the GA, in PSO the particles are adjusted toward the best individual experience (PBEST) and the best social experience (GBEST). However, PSO is unlike a GA in that each potential solution, particle is ''flying'' through hyperspace with a velocity. Moreover, the particles and the swarm have memory; in the population of the GA memory does not exist.

Let $x_{j,d}(t)$ and $v_{j,d}(t)$ denote the dth dimensional value of the vector of position and velocity of jth particle in the swarm, respectively, at time t. The PSO model can be expressed as

$$v_{j,d}(t) = v_{j,d}(t-1) + c_1.\varphi_1.(x_{j,d}^* - x_{j,d}(t-1))$$
$$+ c_2.\varphi_2.(x_d^\# - x_{j,d}(t-1)), \tag{7}$$

$$x_{j,d}(t) = x_{j,d}(t-1) + v_{j,d}(t), \tag{8}$$

where $x_{j,d}^*$ (PBEST) denotes the best position of jth particle up to time t-1 and $x_d^\#$ (GBEST) denotes the best position of the whole swarm up to time t-1, φ_1 and φ_2 are random numbers, and c_1 and c_2 represent the individuality and sociality coefficients, respectively.

The population size is first determined, and the velocity and position of each particle are initialized. Each particle moves according to (7) and (8), and the fitness is then calculated. Meanwhile, the best positions of each swarm and particles are recorded. Finally, as the stopping criterion is satisfied, the best position of the swarm is the final solution. The block diagram of PSO is displayed in Fig. 3 and the main steps are given as follows:

1. Set the swarm size. Initialize the velocity and the position of each particle randomly.

2. For each j, evaluate the fitness value of x_j and update the individual best position $x_{j,d}^*$ if better fitness is found.

3. Find the new best position of the whole swarm. Update the swarm best position $x^\#$ if the fitness of the new best position is better than that of the previous swarm.

4. If the stopping criterion is satisfied, then stop.

5. For each particle, update the position and the velocity according (8) and (7). Go to step 2.
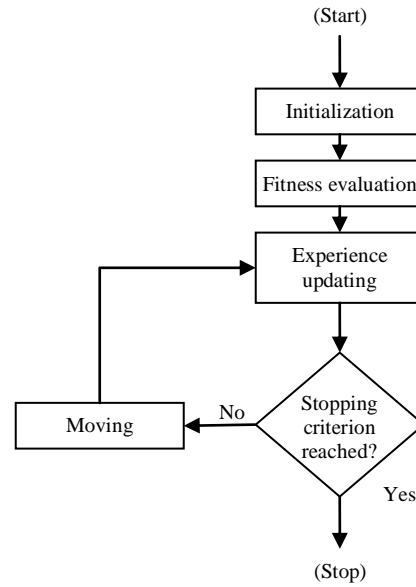
Fig. 3 Block diagram of PSO

IV. PROPOSED METHOD

In the proposed fast fractal encoding using PSO, we reduce the encoding time by reducing the searching time to find a best match domain block for the given range block from all domain blocks.

Flowchart of the fractal encoding of the proposed method is shown in fig. 4.
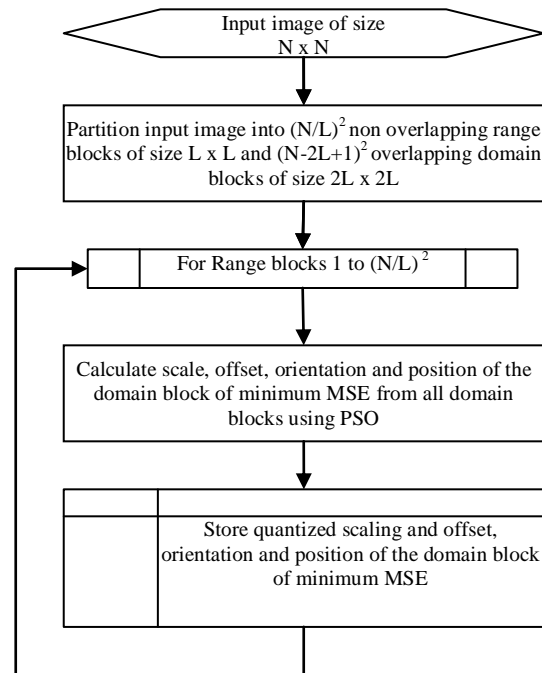
Fig.4 Fractal encoding of proposed method

Domain block of minimum MSE is found by PSO using the steps given below:

1. Set the swarm size proportional to (N-2L+1)^2/(maximum no. of iterations for PSO) and initialize the velocity and the position of each particle randomly

2. Fitness value includes finding MSE between domain block specified by the particles position and given range block using eqn. (3)

3. Update swarm best position if the fitness of the new best position is better than that of the previous swarm.

4. If swarm best position is not changed for some percentage of maximum iteration for PSO, then stop

5. The best position of the particles is updated using eqn. (7) and (8) and goto step 2.

V. XPERIMENTAL RESULTS

The fast fractal encoding using PSO results have been compared to the full search FIC mentioned in the previous sections in terms of encoding time and PSNR.

Fig. 5 shows the original image Lena 256 x 256 at 8bpp. Fig. 6 shows the decoded Lena image using full search FIC and fast fractal encoding using PSO.

The numeric results containing bitrate, encoding time and PSNR of decoded image for various images are tabulated in table I.

Fig. 7 shows the variation in PSNR by varying the stopping criterion in fast fractal encoding using PSO by changing the percentage of maximum iteration of PSO. From fig.7 variation of PSNR with variation of stopping criterion is very less. Hence a 10% of maximum iteration for PSO is choose as stopping criterion.
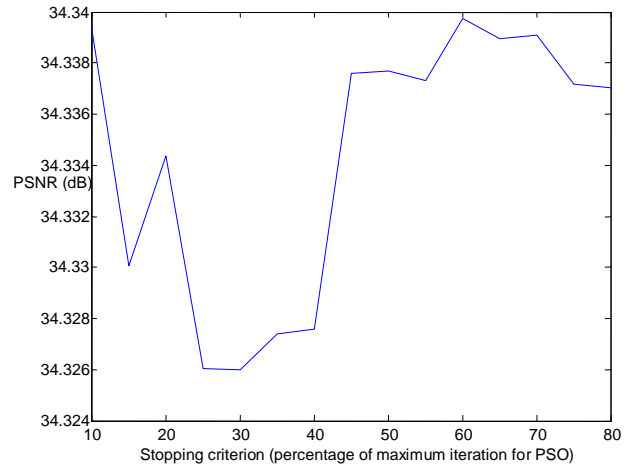
Fig. 7 variation in PSNR by varying the stopping criterion

Fig. 4. Original 256 x 256 Lena image

Fig. 4. Original 256 x 256 Lena image

(a)   (b)

Fig. 5. Decoded Lena image at 0.5 bits per pixel for (a) Full search FIC (b) proposed fast fractal encodind using PSO

(a)   (b)

Fig. 5. Decoded Lena image at 0.5 bits per pixel for (a) Full search FIC (b) proposed fast fractal encodind using PSO

Fig. 4. Original 256 x 256 Lena image

(a)   (b)

Fig. 5. Decoded Lena image at 0.5 bits per pixel for (a) Full search FIC (b) proposed fast fractal encodind using PSO

TABLE I
NUMERIC RESULTS

Input image	Bitrate (bpp)	Method	Encoding Time (hh:mm:ss)	PSNR (dB)
Lena	0.5	Full search FIC	09:07:20	35.80
		Proposed method	00:15:34	35.03
Goldhill	0.5	Full search FIC	09:02:12	33.64
		Proposed method	00:17:37	32.77
Camera man	0.5	Full search FIC	09:02:49	35.11
		Proposed method	00:15:24	34.23

VI. CONCLUSION

Fractal image compression can produce better compression ratio at acceptable quality. By using PSO for fractal coding we can reduce the encoding time with 1.2dB loss in image quality.

VII. REFERENCES

[1] M.F. BARNSLEY, S. DEMKO, ITERATED FUNCTION SYSTEMS AND THE GLOBAL CONSTRUCTION OF FRACTALS, PROC. ROY. SOC. LOND A399 (1985) 243–275.

[2] A.E. Jacquin, Fractal image coding: a review, Proc. IEEE 10 (1993)1451–1465.

[3] A.E. JACQUIN, IMAGE CODING BASED ON A FRACTAL THEORY OF ITERATED CONTRACTIVE IMAGE TRANSFORMATIONS, IEEE TRANSACTIONS ON SIGNAL PROCESSING 1 (1992) 18–30.

[4] M.F. Barnsley, A.D. Sloan, A better way to compress images, BYTE Magazine (1988) 215–233

[5] M. POLVERE, M. NAPPI, SPEED-UP IN FRACTAL IMAGE CODING: COMPARISON OF METHODS, IEEE TRANSACTIONS ON IMAGE PROCESSING 9 (2000) 1002–1009.

[6] T.K. TRUONG, J.H. JENG, I.S. REED, P.C. LEE, A.Q. LI, A FAST ENCODING ALGORITHM FOR FRACTAL IMAGE COMPRESSION USING THE DCT INNER PRODUCT, IEEE TRANSACTIONS ON IMAGE PROCESSING 9 (4) (2000) 529–535.

[7] M.S. WU, J.H. JENG, J.G. HSIEH, SCHEMA GENETIC ALGORITHM FOR FRACTAL IMAGE COMPRESSION, ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE 20 (2007) 531–538.

[8] M.S. WU, W.C. TENG, J.H. JENG, J.G. HSIEH, SPATIAL CORRELATION GENETIC ALGORITHM FOR FRACTAL IMAGE COMPRESSION, CHAOS SOLITONS & FRACTALS 28 (2) (2006) 497–510.

[9] Y. FISHER, *FRACTAL IMAGE COMPRESSION—THEORY AND APPLICATION.* NEW YORK: SPRINGER-VERLAG, 1994.

[10] J. KENNEDY, R.C. EBERHART, PARTICLE SWARM OPTIMIZATION, IN: PROCEEDINGS OF IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS, PERTH, AUSTRALIA, VOL. 4, 1995, PP. 1942–1948.

[11] R.C. EBERHART, J. KENNEDY, A NEW OPTIMIZER USING PARTICLE SWARM THEORY, IN: PROCEEDINGS OF IEEE INTERNATIONAL SYMPOSIUM ON MICRO MACHINE AND HUMAN SCIENCE, NAGOYA, JAPAN, 1995, PP. 39–43.

[12] B. BANI-EQBAL, SPEEDING UP FRACTAL IMAGE COMPRESSION, PROC. SPIE: STILL-IMAGE COMPRESS. 2418 (1995) 67–74.

[13] B. HURTGEN, C. STILLER, FAST HIERARCHICAL CODEBOOK SEARCH FOR FRACTAL CODING STILL IMAGES, SPIE VISUAL COMMUN. PACS MED. APPL. (1993) 397–408.

[14] C. HUFNAGL, A. UHL, ALGORITHMS FOR FRACTAL IMAGE COMPRESSION ON MASSIVELY PARALLEL SIMD ARRAYS, REAL-TIME IMAGING 6 (2000) 267–281.

[15] D. VIDYA, R. PARTHASARATHY, T.C. BINA, N.G. SWAROOPA, ARCHITECTURE FOR FRACTAL IMAGE COMPRESSION, J. SYST. ARCHIT. 46 (2000) 1275–1291.

[16] Y. FISHER, FRACTAL IMAGE COMPRESSION, SIGGRAPH'92 COURSE NOTES 12 (1992) 7.1–7.19.

A.Muruganandham obtained her B.E. degree in 1993 in from Madras University and M.E. degree in 2000 from Bharathithasan University and doing Ph.D in Anna University, Coimbatorei. His areas of interest are Image processing, He is a life member of ISTE and member of IEEE. He is currently working as a Asst. Professor and Electronics and Communication Engineering Sona College of Technology, Salem-05, Tamilnadu, India and has a teaching experience of 15 years. Authored national and international conferences and journals.

Dr.R.S.D.Wahidabanu obtained her B.E. degree in 1981 and M.E. degree in 1984 from Madras University and Ph.D in 1998 from Anna University, Chennai. Her areas of interest are Pattern Recognition, Application of ANN for Image Processing, Network Security, Knowledge Management and Grid Computing. She is a life member of IE, ISTE, SSI, CSI India and ISOC and IAENG. She is currently working as a Professor and Head of Electronics and Communication Engineering, Government College of Engineering, Salem, Tamilnadu, India and has a teaching experience of 27 years. Authored and coauthored 180 research journals in national and international conferences and journals.

Embedding Expert Knowledge to Hybrid Bio-Inspired Techniques- An Adaptive Strategy Towards Focussed Land Cover Feature Extraction

Lavika Goel

M.E. (Masters) Student,
Computer Engineering Department,
Delhi Technological University
(Formerly Delhi College of Engg.),
New Delhi,
India.
Email id - goel_lavika@yahoo.co.in

Dr. V.K. Panchal

Add. Director,
Scientist'G',
Defence Terrain & Research Lab,
DRDO, MetCalfe House
New Delhi
India.
vkpans@ieee.org

Dr. Daya Gupta

Head of Department,
Computer Engineering Department,
Delhi Technological University
(Formerly Delhi College of Engg.),
New Delhi,
India.
dgupta@dce.ac.in

Abstract---The findings of recent studies are showing strong evidence to the fact that some aspects of biogeography can be adaptively applied to solve specific problems in science and engineering. This paper presents a hybrid biologically inspired technique called the ACO2/PSO/BBO (Ant Colony Optimization2/ Particle Swarm Optimization / Biogeography Based Optimization) Technique that can be adapted according to the database of expert knowledge for a more focussed satellite image classification. The hybrid classifier explores the adaptive nature of Biogeography Based Optimization technique and therefore is flexible enough to classify a particular land cover feature more efficiently than others based on the 7-band image data and hence can be adapted according to the application. The paper also presents a comparative study of the proposed classifier and the other recent soft computing classifiers such as ACO, Hybrid Particle Swarm Optimization – cAntMiner (PSO-ACO2), Hybrid ACO-BBO Classifier, Fuzzy sets, Rough-Fuzzy Tie up and the Semantic Web Based classifiers with the traditional probabilistic classifiers such as Minimum Distance to Mean Classifier (MDMC) and the Maximum Likelihood Classifier (MLC). The proposed algorithm has been applied to the 7-band cartoset satellite image of size 472 X 576 of the Alwar area in Rajasthan since it contains a variety of land cover features. The algorithm has been verified on water pixels on which it shows the maximum achievable efficiency i.e. 100%. The accuracy of the results have been checked by obtaining the error matrix and KHAT statistics .The results show that highly accurate land cover features can be extracted effectively when the proposed algorithm is applied to the 7-Band Image , with an overall Kappa coefficient of 0.982.

***Keywords:-** Biogeography based Optimization, Rough Set Theory, Remote Sensing, Feature Extraction, ,Particle Swarm Optimization, Ant Colony Optimization, Flexible Classifier, Kappa Coefficient.*

I. INTRODUCTION

Biogeography is a study of geographical distribution of biological organisms. Species keep changing their geographic location, mostly because of disturbance in ecosystem of their habitat (like drought situations, food adversaries, predators, disease etc). This is mostly a group behavior. They move from an unsuitable habitat to another till a suitable habitat is found. Studying this process gives us the way nature optimizes itself. Various engineers and scientists have and are still working on these nature given algorithms. Various concepts of Particle Swarm Optimization [9], Ant Colony Optimization [11], Evolutionary algorithms are working examples of these nature inspired algorithms. Very recently the concept of Biogeography Based Optimization (BBO) has been introduced in this category.

Biogeography is nature's way of distributing species, and is analogous to general problem solutions. In this algorithm, the optimization is done based on migration of species. It uses the well known procedure that nature uses to balance itself. Every node is given intelligence

to realize whether the resident place is good for it and option to migrate. BBO algorithm is basically used to find the optimal solution to a problem. But satellite image classification is a clustering problem that requires each class to be extracted as a cluster. The original BBO algorithm does not have the inbuilt property of clustering. To extract features from the image, a modified BBO algorithm is used to make the clusters of different features present in the image [3]. Our proposed Algorithm combines the strengths of this modified BBO technique with the hybrid ACO2/PSO Technique for a more refined image classification. The algorithm is also capable of adapting itself to classify a particular land cover feature better than others based on the expert knowledge.

The organization of the paper is as follows: The paper is divided into 7 sections. S*ection 2* presents a brief review on BBO and hybrid ACO2/PSO Techniques .*Section 3* presents the proposed Framework of the Hybrid ACO2-PSO-BBO Algorithm -the dataset used, proposed architecture, and the parameters used. *Section 4* assesses the accuracy of the Proposed Algorithm by analyzing the KHAT Statistics. *Section 5* presents the classification results of the Alwar Image in Rajasthan using ACO2/PSO/BBO Technique and compares its efficiency with the BBO Technique as well as the traditional probabilistic classifiers. *Section 6* presents the classified images using other recent Soft Computing Techniques and provides a comparison of the Soft Computing Classifiers v/s Probabilistic Classifiers. *Section 7* presents Conclusion & future scope of the proposed work.

II. A BRIEF REVIEW OF BBO AND HYBRID ACO2/PSO TECHNIQUES

A. Biogeography Based Optimization

Biogeography Based Optimization is a population based evolutionary algorithm (EA) motivated by the migration mechanisms of ecosystems. It is based on the mathematics of biogeography. In BBO, problem solutions are represented as islands, and the sharing of features between solutions is represented as emigration and immigration. The idea of BBO was first presented in December 2008 by D. Simon[2]. It is an example of natural process that can be modeled to solve general optimization problems. One characteristic of BBO is that the original population is not discarded after each generation, it is rather modified by migration. Also for each generation, BBO uses the fitness of each solution to determine its emigration and immigration rate [2] [1]. In a way, we can say that BBO is an application of biogeography to EAs. In BBO, each individual is considered as a habitat with a habitat suitability index (HSI) [2] [1], which is similar to the fitness of EAs, to measure the individual. Also, an SIV (suitability index variable) which characterizes he habitability of an

island is used. A good solution is analogous to an island with a high HSI, and a poor solution indicates an island with a low HSI. High HSI solutions tend to share their features with low HSI solutions. Low HSI solutions accept a lot of new features from high HSI solutions [1].
.

B. Hybrid ACO2/PSO Optimization

The modified hybrid PSO-ACO for extracting Classification rules given by Nicholas and Frietas [6] uses sequential covering approach for rule extraction [10] which directly deals with both the continuous and nominal attribute-values [9].The new version given by Nicholas and Freitas can be understand as follows-

1. Initially RuleSet is empty(_)
2. For Each class of cases Trs = {All training cases}
3. While (Number of uncovered training cases of class A > Maximum uncovered cases per class)
4. Run the PSO/ACO algorithm for finding best nominal rule
5. Run the standard PSO algorithm to add continuous terms to Rule, and return the best discovered rule BestRule
6. Prune the discovered BestRule
7. RuleSet = RuleSet [BestRule]
8. Trs = Trs – {training cases correctly covered by discovered rule}
 9. End of while loop
10. End of for lop
11. Order these rules in RuleSet by descending Quality

It is necessary to estimate the quality of every candidate rule (decoded particle). A measure must be used in the training phase in an attempt to estimate how well a rule will perform in the testing phase. Given such a measure it becomes possible to optimize a rule's quality (the fitness function) in the training phase and this is the aim of the PSO/ACO2 algorithm. In PSO/ACO [4] the Quality measure used was Sensitivity * Specificity [4]. Where TP, FN, FP and TN are, respectively, the number of true positives, false negatives, false positives and true negatives associated with the rule [4] [8].

Sensitivity Specificity = TP / (TP + FN) TN / (TN + FP)

Equation 1: Original Quality Measure [7]

Later it is modified as follows-

Sensitivity Precision = TP / (TP + F7) TP / (TP + FP)

Equation 2: Quality Measure on Minority Class [7]

This is also modified with using Laplace correction as;

Precision = 1 + TP / (1+ k + TP + FP)

Equation 3: New Quality Measure on Minority Class [7]

Where 'k' is the number of classes.

So, PSO/ACO1 attempted to optimize both the continuous and nominal attributes present in a rule antecedent at the same time, whereas PSO/ACO2 takes the best nominal rule built by PSO/ACO2 and then

attempts to add continuous attributes using a standard PSO algorithm.

III. PROPOSED FRAMEWORK FOR THE HYBRID ACO2/PSO/BBO TECHNIQUE FOR LAND COVER FEATURE EXTRACTION

A. Dataset used

Our objective is to use the proposed hybrid algorithm as an efficient Land cover classifier for satellite image. We have taken a multi-spectral, multi resolution and multi-sensor image of size 472 X 576 of Alwar area in Rajasthan. The satellite image for 7different bands is taken. These bands are Red, Green, Near Infra Red (NIR), Middle Infra Red (MIR), Radarsat-1 (RS1), Radarsat-2 (RS2), and Digital Elevation Model (DEM). The ground resolution of these images is 23.5m and is taken from LISS (Linear Imaging Self Scanning Sensor)-III, sensor. The 7-Band Satellite Image of Alwar area in Rajasthan is given in figure 1.
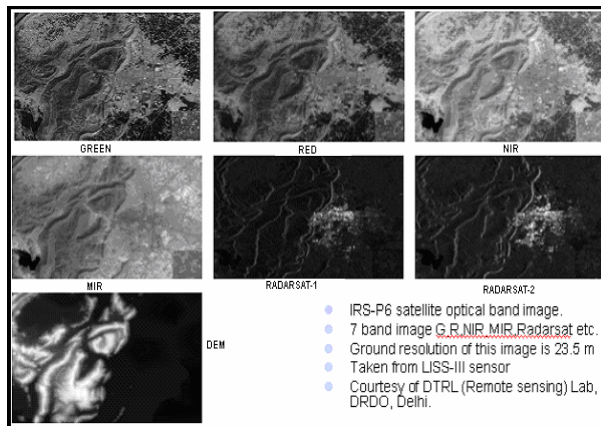
Fig 1. 7-Band Satellite Image of Alwar Area in   Rajasthan

B. Defining Parameters for the Biogeography Based Land Cover Feature Extraction Algorithm

The BBO parameters of the Biogeography Based Land Cover Feature Extraction algorithm are defined as follows [3]:

Definition 1: Each of the multi-spectral bands of image represents one Suitability Index Variable (SIV) of the habitat. Thus, SIV$\in$ C is an integer and C $\in$ [0,255].

Definition 2: A habitat H $\in$ SIV^m where m=7.

Definition 3: Initially there exists a universal habitat that contains all the species to be migrated. Also there are as many other habitats as the number of classes to be found from the image. So the ecosystem H^6 is a group of 6 habitats (one universal habitat and five feature habitats) since 5 features i.e. rocky , barren, water, urban and vegetation are to be extracted from the Alwar Image.

Definition 4: Rough set theory was used to obtain the random clusters of pixels (by using discretization and partitioning concept of rough set theory) and each of the resulting cluster will be considered as mixed species that migrate from one habitat to another. These species can also be termed as 'elementary classes' of a habitat.

Definition 5: Standard deviation of pixels is used as Habitat Suitability Index to help in image classification.

Definition 6: The original BBO algorithm proposed the migration of SIV values from a high HSI habitat to a low HSI habitat. In the above algorithm, rather than moving SIV, the species are moved altogether from a universal habitat to feature habitat. The species do not remain shared: it is removed from the universal habitat and migrated to the feature habitat.

Definition 7: Maximum Immigration rate and Maximum Emigration Rate are same and equal to number of species in the habitat. [2] Maximum species count (Smax) and the maximum migration rates are relative quantities.

Definition 8: Since mutation is not an essential feature of BBO, it is not required in the proposed algorithm. Elitism, too, is an optional parameter; it has not been in the modified BBO Algorithm.

C. Proposed Architecture

The process of Biogeography Based Land Cover Feature Extraction is divided into three steps:

➢ The first step considers a class and concatenates it with various training sets (i.e. water, vegetation, rocky, barren and urban). These classes and training sets are saved as excel sheets containing x coordinate, y-coordinate, DN values of all the bands. After concatenation each result is stored in a different sheet.

➢ The next step is to use a Heuristic procedure to decide which land cover property each class belongs to. This is done (in Matlab [13]) by comparing the mean of the Standard Deviation for each of these classes (defined as the Fitness Function) with the Standard Deviation of the Feature Habitat class, using a specific threshold value [3].

➢ Therefore, Fitness function = difference of the mean of the Standard Deviation for each of these classes. Feature Habitat class = class which contains the standard training set pixels of the 7-Band Image of the Alwar region for comparison.

➢ In the final step, this function decides which value of mean of standard deviation has minimum difference from the original class.
i.e.

HSI = Standard Deviation for each of the classes

➢ If this value is within the threshold then that class (species) will migrate to that habitat.[3] If not it can migrate to other class .This can be mathematically represented as below –

Let x_i represent one of the 20 Rosetta [12] classified rough set classes i.e. the universal habitat and y_i training set gray level values i.e. the feature habitat for the ith band of the 7-band image for each of the 5 land cover features to be extracted,

Then,

If $| [\Sigma\sigma_{xi}/n] - [\Sigma\sigma_{yi}/n] |^6_{j=1} <$ threshold ,
　　　　UH　　　　FH

　　where,

　　　　UH=Universal Habitat
　　　　FH=Feature Habitat

➢ 　then the feature is decided as 'j' i.e. the said Equivalence class corresponds to the feature 'j'.
　　else
　$j =1$ i.e. it is treated as unclassified .

If it belongs to no class it can simply move to the universal habitat and divides itself to a number of classes which then choose their habitats .The BBO approach can handle a little of inaccuracy in training sets. BBO also takes up inaccurate classes and tune it up for better results.

In this paper we have implemented an integration of Biogeography based land cover feature extraction with the ACO2/PSO technique for features extraction from a satellite image. The proposed architecture of our hybrid algorithm is as follows-

➢ 　The image used is the 7-Band Satellite Image of size *472 X 576* of the Alwar Region in Rajasthan. The satellite image is divided into 20 clusters.

➢ 　We use rough set theory toolkit i.e. *Rosetta software* [12] for dicretizing each of the 20 clusters using the semi-naïve Algorithm & then partition each of them based on the band which is able to classify the particular feature that we want to extract from the image. Depending on our application, for example, if we want to extract the barren area more efficiently, we choose the green band and for rocky region extraction, we choose the MIR Band. The RS-1 and RS-2 bands

are used to extract the urban area and also for extracting the edges of rocky region from the 7-band image. However, the drainages of rocky region are best viewed in the Red band and water and vegetation pixels are best viewed in NIR and MIR Bands. For our illustration, we choose the NIR and MIR band of the 7-band image since we want to extract the water pixels effectively and clearly identify the water body in the image and these are the bands in which the water feature is particularly more highlighted and best viewed. Therefore, we use the NIR and the MIR bands for discretization and partitioning step in the semi – naïve algorithm used for creating rough set equivalence classes, thus creating Equivalence classes for each of the clusters. This is what is termed as Unsupervised Classification .Each of these resultant classes are put in the Universal Habitat.

➢ 　Based on the results obtained on applying the BBO algorithm to the 7-Band Image of Alwar region for Land Cover Feature Extraction , we observe that we are able to classify some particular feature's pixels (in our case ,water) with greater efficiency than the other features based on the band chosen & hence, we apply BBO Technique on that particular cluster of the Satellite image of Alwar region since this is the cluster which gives the maximum classification efficiency because it predominantly shows the presence of the feature that is most efficiently classified by the BBO Algorithm.

➢ 　We then apply ACO2/PSO Technique [4] on the remainder of the clusters of the image by taking the training set for the 7-Band Alwar image in .arff [4] format as input to generate rules from it using the open source Tool [4] and then applying them on each of the remainder clusters checking for pixel validation for each pixel in the cluster & thus obtain a refined classification of the image .

➢ 　Therefore, the working of our proposed hybrid algorithm can be summarized in the form of the following equation and mathematically explained as follows-

where Universal Habitat contains the rough set classified equivalence classes and the feature habitat consists of the expert generated training set of the original Alwar image in 7-bands.
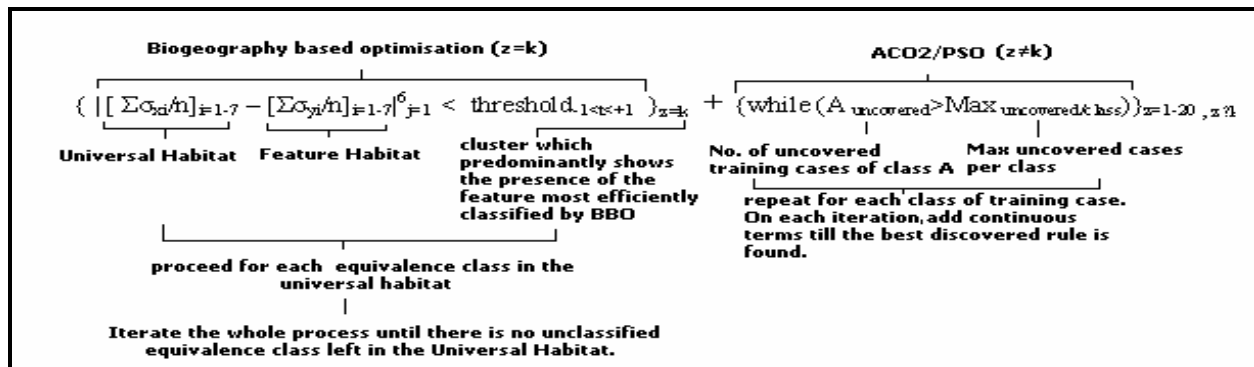
➢ Then, for z=k , we proceed in the following manner for the BBO Optimizer -i.e. for each ith band where 'i' ranges from 1-7 , we calculate the difference in the standard deviation of the ith band of the Universal Habitat and the ith band of the Feature Habitat containing the expert generated training set of the image. If this difference is the minimum for the feature 'j' and also less than the pre-specified threshold value of ,-1 < t < +1,then that particular equivalence class is classified as the feature 'j' else j=1(unclassified).The process is repeated for each equivalence class until there is no equivalence class left in the universal habitat and the whole process is iterated till there is no unclassified Equivalence class left.

➢ For z=1-20, where z ≠ k, we use the ACO2/PSO Optimization, wherein the training set for the 7-Band Alwar image in .arff [4] format is used as input to generate rules from it using the open source Tool [4] for each class of training case and on each iteration, we add continuous terms till the best discovered rule is found. The classification rules are then applied on the remainder of the clusters checking for pixel validation on each of them.

➢ Hence, we obtain a more refined classified image with an improved Kappa coefficient which is much better than the Kappa Coefficient we get when we apply the original BBO Algorithm on the 7-Band Image.

This in turn leads us to the improved flexible Hybrid version of the BBO Algorithm for Satellite Image Classification which will classify the particular feature chosen by the band used in the unsupervised classification , most efficiently ,which is in turn based on the expert knowledge and the band information contained in the training set of the particular area. Thus, we have efficiently exploited the properties of the BBO technique to adapt itself to

a more focussed classification which upon integrating with the ACO2/PSO Technique makes an advanced classifier. Hence, we have obtained a hybrid algorithm which can be adapted to incorporate the expert knowledge for a more flexible, efficient and refined classification. The proposed overall Architecture of this Hybrid ACO2/PSO/BBO Technique is illustrated by means of a flowchart in fig. 2.

IV. ACCURACY ASSESSMENT OF THE PROPOSED ALGORITHM

Accuracy assessment is an important step in the classification process. The goal is to quantitatively determine how effectively pixels were grouped into the correct feature classes in the area under investigation.

Fig. 3 shows the data distribution graph plotted between the average of the Standard Deviations of each land cover feature viz water, urban, rocky, vegetation and barren (plotted on the y-axis) for each of the 7-Bands of the image i.e. Red, Green, NIR, MIR ,RS1, RS2 and DEM (plotted as the x-axis) .From the graph, it can be observed that the minimum difference between the average standard deviations of the NIR and the MIR bands of the Alwar Image is achieved in particularly two land cover features , those of water and urban area ,both of which exhibit the same graph pattern in the NIR and the MIR bands .
i.e.
| average of standard deviation of NIR band ~ average of standard deviation of the MIR band | $_{lowest}$ = {water, urban}

Hence, it can be concluded that these are the two features that will be most efficiently classified by our hybrid algorithm which works in the NIR and MIR bands .
Now we proceed to calculate the classification accuracy of our proposed algorithm using the classification *error matrix*. Error matrices compare, on category-by category basis, the relationship between known reference data

(ground truth) and the corresponding results of an automated classification. We took 150 vegetation pixels, 190 Urban pixels, 200 Rocky pixels, 70 water pixels, 170 barren pixels from the training set and the error matrix obtained is shown in Table II.

The error matrix's interpretation along column suggests how many pixels are classified correctly by algorithm. The diagonal elements (diagonal elements indicate the no. of correctly classified pixels in that category) . From Table I (simple BBO Classifier) , it is evident that the BBO Technique shows the maximum efficiency on the water pixels since it

classifies 69 out of 70 pixels correctly as water pixels with only 1 omission error wherein it classifies 1 pixel as rocky one. However, BBO is not an efficient classifier for the urban feature which is also evident from Table II, wherein whole 190 out of 190 pixels were correctly classified as Urban pixels whereas simple BBO Classifier in table I could only classify 88 pixels correctly as urban pixels and it classified 91 pixels wrongly as barren ones. Therefore, we use the Hybrid Technique to classify , in particular the Water and the Urban pixels, with almost 100% efficiency (with no
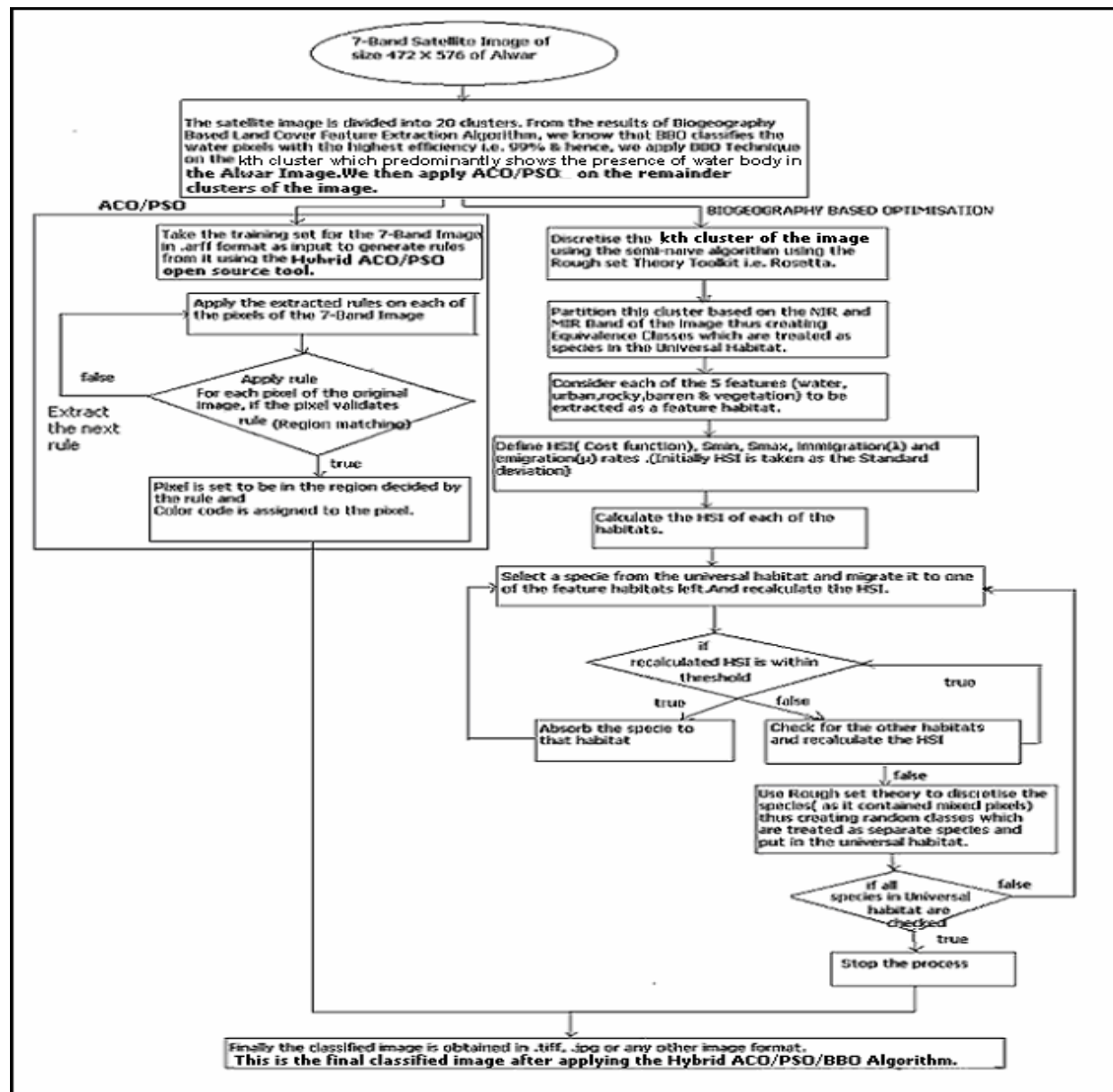

Fig 3.Overall Framework of the hybrid ACO/BBO/PSO Algorithm

omission errors) ,since for water pixels, we achieve zero omission and commission error (ideal classification) through our algorithm and for urban pixels, a commission error of just 5 in 195 with no omission error (near-ideal classification). This is what was also reflected earlier, from the data distribution graph plotted .

The Kappa coefficient of the Alwar image is calculated using the method described Lillesand and Kiefer. The Kappa (K) coefficient of the Alwar image is 0.9818 which indicates that an observed classification is 98.82% better than one resulting from chance.
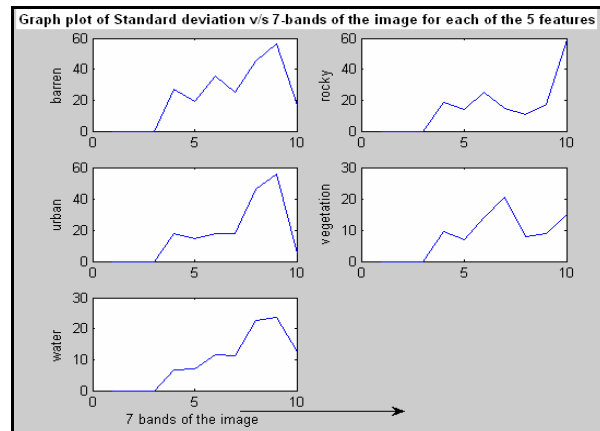

Figure 3. Graph plot of the Standard Deviations of each Land Cover feature v/s each of the 7-Bands in which the Alwar Image is viewed.

Table I. Error matrix when only BBO is applied
Kappa coefficient = 0.6715

	Vegetation	Urban	Rocky	Water	Barren	Total
Vegetation	127	9	0	0	2	138
Urban	0	88	1	0	32	121
Rocky	6	2	176	1	17	202
Water	0	0	3	69	0	72
Barren	17	91	20	0	119	247
Total	150	190	200	70	170	780

Table II. Error Matrix when Hybrid ACO2/PSO-BBO technique is applied.
Kappa Coefficient=0.9818

	Vegetation	Urban	Rocky	Water	Barren	Total
Vegetation	142	0	0	0	0	142
Urban	5	190	0	0	0	195
Rocky	0	0	198	0	3	201
Water	0	0	0	70	0	70
Barren	2	0	1	0	163	166
Total	149	190	199	70	166	774

V. RESULTS AND DISCUSSION

 Based on the results obtained on applying the BBO algorithm to the 7-Band Image of Alwar region for Land Cover Feature Extraction, we observe that are able to classify water pixels with the highest efficiency i.e. 99% efficiency and these are the pixels best viewed in the NIR and MIR bands in the BBO Technique & hence, we apply BBO Technique on the 16th cluster of the Satellite image of Alwar region (z=16) since this is the cluster which predominantly shows presence of water body in the Alwar Image . However, BBO shows poor efficiency, in fact the poorest, in classifying the urban pixels as shown in fig. 4. Here the encircled region in the BBO Classified Image shows that BBO wrongly classifies the urban pixels as barren ones which is also reflected from Table I where BBO classifies 91 urban pixels wrongly out of 190 total urban pixels.

Therefore, in order to classify the urban pixels efficiently, we then apply ACO2/PSO Technique [4] on the remainder of the clusters of the image (z $\neq$16) by taking the training set for the 7-Band Alwar image in .arff [4] format as input to generate rules from it using the open source Tool [4] and then applying them on the remainder of the clusters checking for pixel validation for each pixel in the cluster & thus obtain a more refined classification of the image with an improved Kappa coefficient of 0.9818 which is much better than the Kappa Coefficient of 0.6715 [3] we get, when we apply the original BBO Algorithm on the 7-Band Image . This in turn leads us to the improved Hybrid version of the BBO Algorithm for Satellite Image Classification where both the urban and the water features are classified with the highest efficiency i.e. almost 100% with no omission errors followed by rocky with only 1 omission error (column wise error) and thereafter barren and vegetation features ,respectively. After applying the proposed algorithm to the 7-band of Alwar Image, the classified image is obtained in figure 5. From the figure, it is clearly shown that our proposed ACO2/PSO-BBO classifier is able to correctly classify the encircled region as urban which was wrongly classified by the simple BBO Classifier. The yellow, black, blue, green, red color represents rocky, barren, water, vegetation, urban region respectively. As the threshold limit of HSI matching is lowered, the species do not get absorbed in the feature habitat and return to universal habitat. Those species are further discretized and classified in next iterations (generation).

From the figures 4 & 5, it is evident that the Hybrid ACO2/PSO-BBO Technique produces a more refined image as compared to the BBO classified image. Figure 6 compares the Hybrid ACO2/PSO-BBO Technique with the Minimum Distance Classifier (MDC) & Maximum Likelihood Classifier (MLC). A comparison of the Kappa Coefficients of the Hybrid ACO2/PSO/BBO Classifier with the Traditional Classifiers is given in Table III.
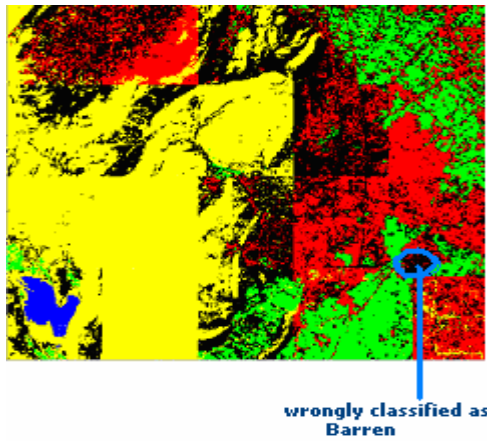
Fig.4. Classified image after applying BBO
(with Kappa Coefficient=0.6715)
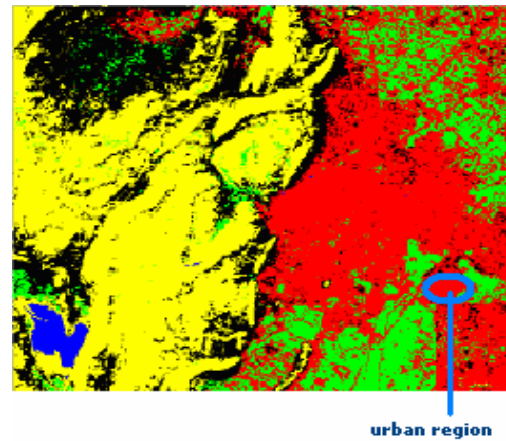
Fig 5. Hybrid ACO2/PSO/BBO Classified Image
(Kappa Coefficient=0.98182)
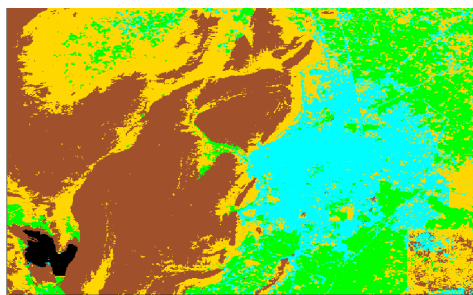
Minimum Distance Classifier
(Kappa Coefficient=0.7364)
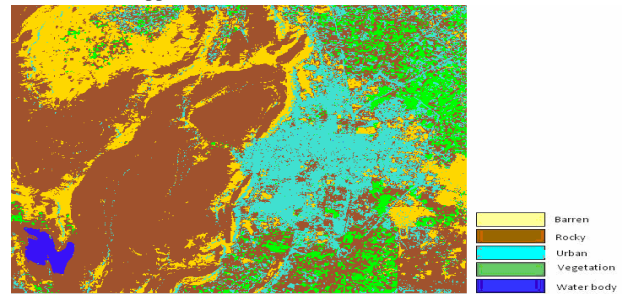
Maximum Likelihood Classifier
(Kappa Coefficient=0.7525)

Fig 6. A comparison with the Traditional Probabilistic Classifiers

Table III. A comparison of Hybrid ACO2/PSO-BBO Classifier with traditional classifiers.

Minimum Distance Classifier(MDC)	Maximum likelihood Classifier(MLC)	Biogeography Based Optimization (BBO)	Hybrid ACO2/PSO-BBO Classifier
0.7364	0.7525	0.6715	0.98182

VI. CLASSIFICATION RESULTS OF OTHER SOFT COMPUTING TECHNIQUES USED FOR SATELLITE IMAGE CLASSIFICATION

From the above discussion, it is evident that the Hybrid ACO2/PSO/BBO Approach is a much efficient classifier as compared to the traditional probabilistic classifiers such as the MDMC and MLC. However, this Hybrid ACO/PSO/-BBO technique also produces comparable results with the image classification results of the other recent soft computing classifiers as shown below. Fig 7(a) shows the Fuzzy Classification of Alwar region which has a Kappa –Coefficient of 0.9134. Fig 7(b) presents the results of an integrated Rough –Fuzzy Tie Up Approach which has a Kappa Coefficient of 0.9700. Fig 7(c) applies the cAntMiner Algorithm on the Alwar Region which has a Kappa Coefficient of 0.964. Fig 7(d) shows the result of applying the hybrid ACO-BBO Technique on the Alwar Image which has a Kappa-Coefficient of 0.96699. Fig 7(e) applies the

Hybrid ACO2/PSO Classifier which has a Kappa Coefficient of 0.975. Fig 7(f) presents the results of the Semantic Web Based Classifier on the image with a Kappa Coefficient of 0.9881[5]. The Table IV below compares the Kappa Coefficients of the Soft Computing Classifiers v/s the Traditional Probabilistic Classifiers .From the Table, it is clearly reflected that Soft Computing Classifiers are much more refined & efficient than the Probabilistic Classifiers.

VII. CONCLUSION & FUTURE SCOPE

Discrepant uncertainties inherent in satellite remote sensing images for geospatial features classification can be taken care of by use of soft computing techniques effectively. For the purpose, Rough Sets, Fuzzy Sets, Rough-Fuzzy Tie-up, Ant Colony

(a) Fuzzy Classification of Alwar Region
(Kappa Coefficient=0.9134)

(b) Rough-Fuzzy Tie Up
(Kappa Coefficient=0.9700)

(c) cAntMiner Algorithm
(Kappa Coefficient=0.964)

(d)Hybrid ACO-BBO Algorithm
(Kappa Coefficient=0.96699)

(e) Hybrid ACO2/PSO Algorithm
(Kappa Coefficient-0.975)

(f) Semantic Web Based Classifier
(Kappa Coefficient=0.9881)
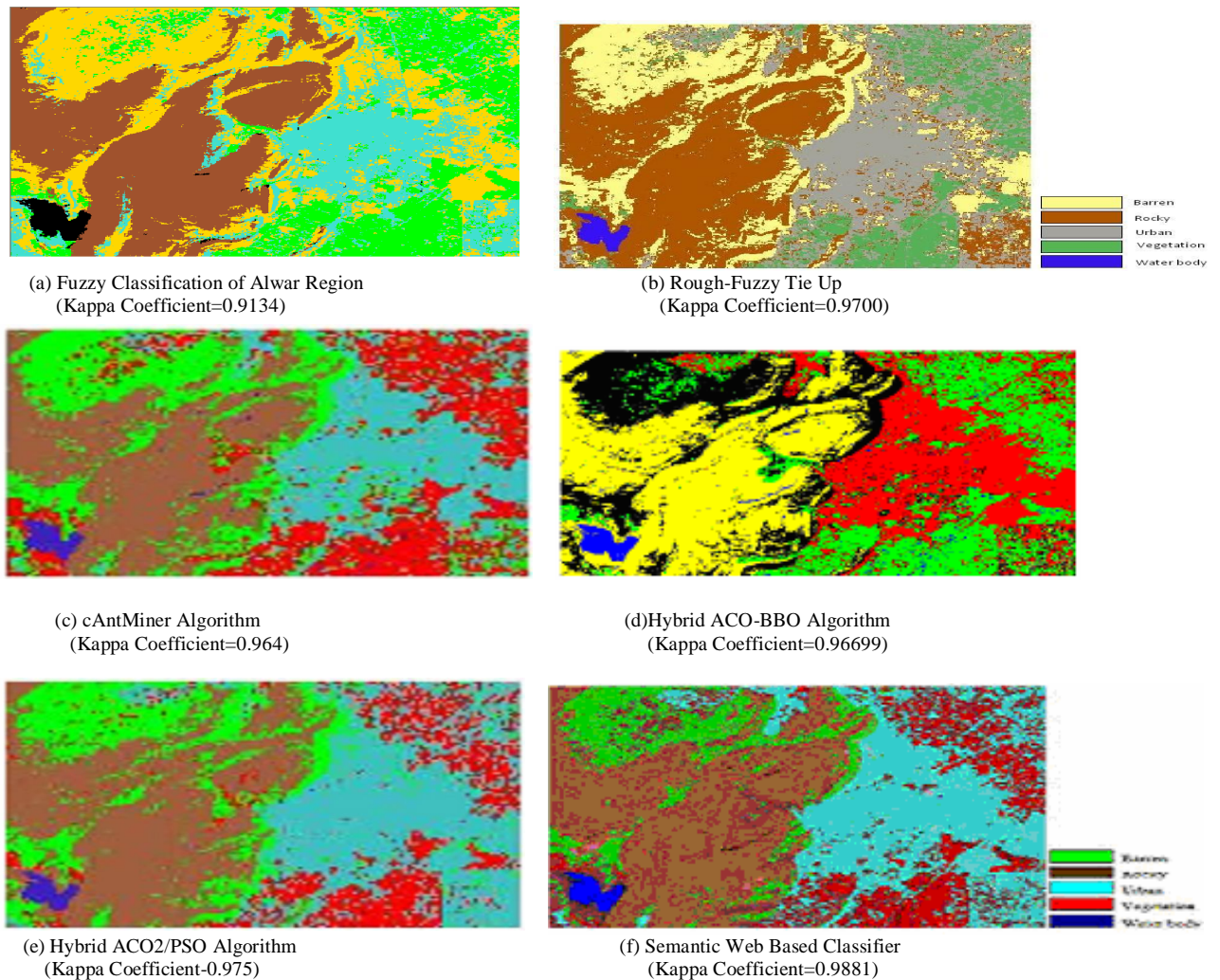
Fig 7. Classified Images of Alwar Region after applying various Soft Computing Techniques

Table IV. Kappa Coefficient (k) of Soft Computing Classifiers v/s Probabilistic Classifiers

Minimum Distance Classifier (MDC)	Maximum Likeli-hood Classifier (MLC)	Fuzzy set	Rough-Fuzzy Tie up	cAnt-Miner	Hybrid ACO2/ PSO	Semantic Web Based Classifier	Biogeo-graphy Based Classifier	Hybrid ACO-BBO Classifier	Hybrid ACO2/ PSO/BBO Classifier
0.7364	0.7525	0.9134	0.9700	0.964	0.975	0.9881	0.6715	0.96699	0.98182

(Probabilistic Classifiers)   (Soft Computing Classifiers)

Technology Growth

Optimization, Particle Swarm Optimization, semantic web-based classification and   Biogeography Based Optimization methods are analyzed in the paper. Semantic-web based image classification is added, as a special instance. Decision system required for any supervised classification can be made consistent and free from indecisive regions by using this spectrum of methods. The Land cover Classification is taken as a case study. It is perceived, from this research, that Kappa coefficient, a well founded metric for assessing

the accuracy of classification in remote sensing community, may be used for comparative study of the results from soft computing methods.

This paper presents a novel approach wherein BBO can be combined with ACO/PSO to solve the Image Classification problems in remote sensing for feature extraction from high resolution multi-spectral satellite images .BBO can be used for further refinement of the image classified by simple ACO algorithms such as the cAntMiner Algorithms ,since BBO refines its

solutions probabilistically after each iteration unlike ACO/PSO which produces new solutions with each iteration and also it is particularly flexible to incorporate the expert knowledge for a more focussed image classification. Hence using a combination of the two techniques i.e. the ACO2/PSO and BBO Technique, can be of major benefit.

In future, the algorithm efficiency can be further improved by lowering the threshold value used in BBO algorithm thus leading to more iterations and refined results. Also, we can further divide the image into more clusters so that a more accurate comparison can be made and the decision about which of the two techniques to be applied on the particular cluster , can be further streamlined. The system performance can be further increased by using better unsupervised classifications and better training sets.

ACKNOWLEDGMENT

REFERENCES

[1] Lavika Goel, V.K. Panchal, Daya Gupta, Rajiv Bhola ,"Hybrid ACO-BBO Approach for predicting the Deployment Strategies of enemy troops in a military Terrain Application" in 4th International MultiConference on Intelligent Systems & Nanotechnology (IISN-2010), February 26-28, 2010.

[2] D.Simon ,"Biogeography-based Optimization" ,   , in IEEE Transactions on Evolutionary Computation, vol. 12, No.6, IEEE Computer Society Press. 702-713., 2008.

[3] V.K. Panchal , Samiksha goel, Mitul Bhatnagar, "Biogeography Based Land Cover Feature Extraction" , in VIII International Conference on Computer Information Systems and Industrial Management (CISIM 2009) ,Coimbatore,December 2009.

[4] Shelly Bansal, Daya Gupta, V.K. Panchal ,Shashi Kumar, "Remote Sensing Image Classification by Improved Swarm Inspired Techniques" in   International Conference on Artificial Intelligence and Pattern Recognition (AIPR-09), Orlando, FL, USA ,July 13-16,2009 .

[5] Sonal Kumar, Daya Gupta, ,V.K.Panchal, Shashi Kumar, "Enabling Web Services For Classification Of Satellite Images", in 2009 International Conference on Semantic Web and Web Services (SWWS'09), Orlando, FL, USA ,July 13-16,2009.

[6] Holden and A.A. Freitas " A hybrid particle swarm/ant colony algorithm for the classification of hierarchical biological data." In: Proc. 2005 IEEE Swarm Intelligence Symposium (SIS-05), pp. 100-107, IEEE, 2005.

[7]. Holden and A.A. Freitas " Hierarchical Classification of GProtein-Coupled Receptors with a PSO/ACO Algorithm" In: Proc. IEEE Swarm Intelligence Symposium (SIS-06), pp. 77-84. IEEE, 2006.

[8] S. Parpinelli, H.S. Lopes and A.A. Freitas "Data Mining with an Ant Colony Optimization Algorithm",   in IEEE Trans. On Evolutionary Computation, special issue on Ant Colony algorithms, 6(4), pp. 321-332, Aug 2002.

[9] Bratton and J. Kennedy "Defining a Standard for Particle Swarm Optimization" in proceedings of the 2007 IEEE Swarm Intelligence Symposium, Honolulu, Hawaii, USA, April 2007.

[10] J. Hand. Wiley "Construction and Assessment of Classification Rules"., 1997.

[11] Dorigo and T. Stuetzle "Ant Colony Optimization" in MIT Press, 2004.

[12] Óhrn, A. and Komorowski, J., ROSSETA " A Rough Set tool kit for analysis of data"   ,in roc.3rd International Joint   Conference   on   information   Sciences,   Vol ,Durham,NC,March 1997.

[13] The MATLAB ver 7, The MathWorks, Inc.

AUTHORS PROFILE

Lavika Goel has done B-Tech (Hons.) in Computer Science & Engineering & scored 78% marks from UP Technical University, Lucknow (India) in 2008 and currently pursuing Master of Engineering in Computer Technology & Applications from Delhi College of Engineering, New Delhi of India ,batch 2008-2010. She is currently working in Defence Terrain & Research Lab at Defence & Research Development Organisation(DRDO) as a trainee for the completion of her final year project.The work done in this paper is also a part of her M.E. Thesis work.

Dr. V.K. Panchal is Add. Director at Defence Terrain Research Lab, New Delhi. Associate Member of IEEE (Computer Society) and Life Member of Indian Society of Remote Sensing.He has done Ph.D in Artificial Intelligence and is currently working as Scientist 'G' at DRDO,Delhi.Chaired sessions & delivered invited talks at many national & international conferences. Research interest are in synthesis of terrain understanding model based on incomplete information set using bio-inspired intelligence and remote sensing.

Dr. Daya Gupta is the Head of Computer Engineering Department, Delhi College of Engineering, New Delhi. She has done M.Sc. (Maths),Post M.Sc. Diploma (Computers Sc.) from IIT, Delhi, Ph.D. She is a Member of CSI and her specialization is in Computer Software. She has chaired many sessions and delivered invited talks at many national and international conferences.

On Multi-Classifier Systems for Network Anomaly Detection and Features Selection

Munif M. Jazzer

Faculty of ITC,
Arab Open University-Kuwait
Kuwait.

Mahmoud Jazzar

Dept. of Computer Science
Birzeit University
Birzeit, Palestine

Aman Jantan

School of Computer Sciences
University of Science Malaysia
Pulau Pinang, Malaysia

Abstract—**Due to the irrelevant patterns and noise of network data, most of network intrusion detection sensors suffer from the false alerts which the sensors produce. This condition gets worse when deploying intrusion detection measures in real-time environment. In addition, most of the existing IDS sensors consider all network packets features. Using all packets features for network intrusion detection will result in lengthy and contaminated intrusion detection. In this research we highlight the necessity of using important features in various anomaly detection cases. The paper presents a new multi-classifier system for intrusion detection. The basic idea is to quantify the causal inference relation to attacks and attacks free data to determine the attack detection and the severity of odd packets. Initially, we have refined the data patterns and attributes to classify the training data and then we have used the SOM clustering method and the fuzzy cognitive maps diagnosis to replicate attacks and normal network connection. Experimental results shows that the classifiers gives better representation of normal and attack connection using significant features.**

Keywords- Anomaly Detection; SOM; FCM; Security

I. INTRODUCTION

The basic function of anomaly-based sensors is to detect any deviation from normal system behavior. However, clear merits between normal and abnormal patterns are very difficult to realize in practice especially when new systems are added or removed from the system network [1, 2]. As a solution, we are trying to tackle this problem by implementing unsupervised learning and knowledge discovery techniques such that there is no need for training the system on clean data.

The typical network-based IDS process system activities based on network data and make a decision to evaluate the probability of action of these data to decide whether these activities are normal or intrusions [1]. In order to evaluate the system activity and trace the probability of action of normal vs. intrusive data, the basic knowledge of network attacks is necessary. The problem is that network attacks may not happen at single action such that one massive attack may start by seemingly innocuous or by small probe action to take place [3]. Such situation articulates the need for a defense-in-depth strategy. At this point, we have considered the domain knowledge of network data, thus we need to extract the causal relation of these data and make inference with it. First, we cleanse the data and then diagnose the clean data patterns.

In this paper, we have used fuzzy cognitive maps (FCM) [4, 5] to express the causal relation of data and calculate the severity and relevance to attacks or normal connection. We have also used the SOM method [6] to help us evaluate the related data patterns and attributes. As a result, benign concepts can be dropped or ignored and other can be addressed as a potential risk of attacks or error caused.

The main objective of this paper is to present a new multi-classifier system based on causal knowledge acquisition and show its effectiveness for anomaly detection. Features selection measures are also considered and illustrated in various detection cases. The detailed system process overview is illustrated in Fig. 3. A brief summary of the exploration modules and its processes details are available in Table II. The rest of the paper is organized as follows: Anomaly detection in network-based IDS and related issues are discussed in section II, the related works are discussed in section VI, the classifiers detection process in section III, and the features selection process in section IV. Section V describes the performance evaluation, related discussion, concluding remakes and future work.

II. ANOMALY DETECTION

A typical anomaly-based detection system works on the notion that abnormal behaviors and activities are different enough from normal (legitimate) behaviors profile.

In anomaly detection, patterns are analyzed based on some measures (statistical, threshold, rule-based ...) to determine the events or activities that are malicious or abnormal. The most attractive thing here is that the IDS that employ these kinds of detection mechanisms have abilities to detect symptoms of attacks without previous knowledge of their attack details which makes them ideal for detecting the newly rising attacks signatures [7]. Furthermore, information produced by anomaly-based detection systems can be used to define signatures for misuse-based detection systems. On the other hand, the output produced from anomaly-based detectors can be in turn used as information source for misuse-based detectors i.e. to double check for legitimate activities that might be intrusion [8]. As result, anomaly detectors are attractive and can play a measure part in the future IDS. A block diagram of a typical anomaly detection system is shown in Figure 1.
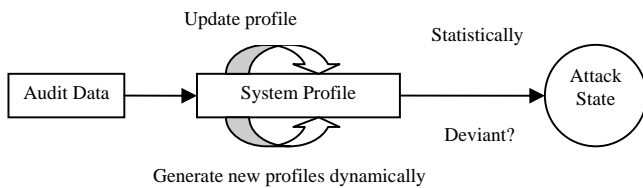
Figure 1. A typical anomaly IDS [7].

The main issues with the anomaly-based detectors are that they produce high number of false alerts [7]. According to [9], anomaly detectors are tending to be computationally expensive. This is because there are several metrics which are often maintained and often need to be updated against every system activity; and they might be gradually trained incorrectly to recognize abnormal behaviors as normal in the long run due to the insufficient data.

In this study, we have assumed that the neural networks behaviors in SOM will learn patterns of the normal system behavior and continually produce profiles to incorporate with the fuzzy logic behaviors in the FCM. This is also to determine the appropriate membership function which will help in reducing false alerts and increasing the detection accuracy of the detection sensor [10].

III. THE CLASSIFIER INTRUSION DETECTION

The ability of detecting/preventing new attacks without prior knowledge of the attack behavior is a tough task, especially the way of determining the input features to monitor normal versus intrusive behavior. For this challenging task, we decide on unsupervised learning techniques as they are the best suited for such situation [27].

The focus here is to provide a multi-classifier system which can work as an inference engine supplement for enhancement of the IDS capability. Using the classifiers system, we can determine the importance of features in various anomaly detection cases.

In order to build the inference engine classifiers system, we have used the unsupervised learning method so-called Kohonen's maps (SOM) [6] for clustering and recognition of input data and the fuzzy cognitive maps (FCM) [5] to detect features relevancy. The FCM use causal reasoning to assess the SOM output and then model the final decision. FCM are ideal causal knowledge acquiring tool with fuzzy signed graphs which can be presented as an associative single layer neural network [4]. Using FCM, our methodology attempt to diagnose and direct network traffic data based on its relevance to attack or normal connections.

By quantifying the causal inference process we can determine the attack detection and the severity of odd packets. As such, packets with low causal relations to attacks can be dropped or ignored and/or packets with high causal relations to attacks are to be highlighted. In the following subsections, we elaborate the classifiers system modules. Figure 3 shows the overall detection process.

A. Preprocessor Module

Data preprocessing module performs the final preparation of the target data records. This includes the slicing of the large dataset. The selection criteria based on pre-user defined mechanisms or threshold value, and the number of the starting row in the given dataset. First, we introduce the input file with all the input vectors then we put the number of vectors required to read, the number of levels and the threshold value. In this module, the user can introduce the number of neurons and the selected features which will be used in each SOM level. After that, the user can train and save the neurons state accordingly for each training level. The elapse time is the difference between the first and the last level according to the user predefined number of levels.
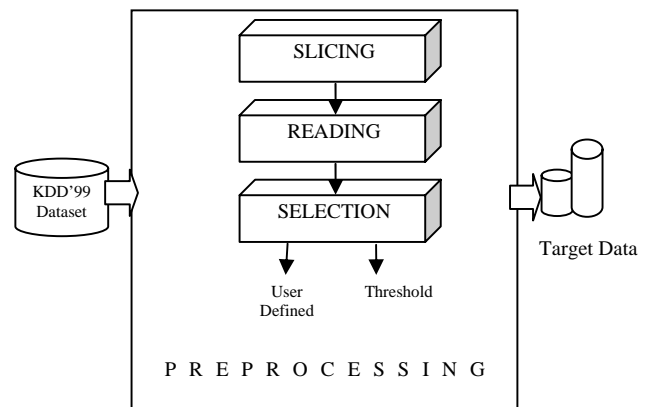
Figure 2. Preprocessing module.

This module involves slicing the dataset into five classes. Each class symbolic-valued features are mapped into numeric-valued features. Symbolic features such as protocol type symbols (TCP, UDP, and ICMP) were mapped into integer values. More details about the data used and the data classes are available in section V. Each symbol data is corresponded to a position in the labels array and this position will be used to fill the input vector. In this module, we have focused on the final preparation of the target data to be presented to the subsequent module.

The prime importance of this module join up by the fact that finding or discovering related patterns in a data set is an instructive process, with slight or even no former knowledge about the structure of the given dataset to be examined [21]. Hence, dependence on clean dataset can give more confidence that the assumption drawn from the pattern exploration output can be treated as being precise to the model of the data being examined. Moreover, the redundant and non related patterns can be dropped earlier to avoid congestion on the subsequent operations. Thus, it gives the system vigilant and the flexibility of features selection for further exploration of attacks details.

B. Data Mining Module

Data mining module is the first important component of the classifiers system. The task of this module is to generate cluster information such that generates logical and homogeneous clusters from the input dataset. To achieve that task, a network

classifier (SOM) is used to do an initial recognition of the network traffic flow to detect abnormal behaviors. To achieve the key objective of the data mining module, the data is first passed through the SOM such that the data and its relevant features are represented by the SOM. The learnt SOM then passed through the fine tuning module for knowledge discovery using the FCM exploration.

Two stages are required in order to create the SOM which are the initialization and the training of the SOM. The initialization process sets up the map with the desired dimensions and initial weights for each unit of the map. The training process allows the map to adapt to the features of the data set during a number of epochs.

At each epoch one input vector x is compared to all neurons weights w with a distance function (Euclidean or Manhattan) to identify the most similar nodes so-called the best matching unit (BMU). Once the BMU has been found, the neighboring neurons and the BMU itself are updated according to the following rule:

$$w_i(t+1) = w_i(t) + h_{ci}(t)[x(t) - w_i(t)] \qquad (1)$$

Where: t is an integer which denotes time, $h_{ci}(t)$ is the neighborhood function around the winner unit c and $x(t)$ is the input vector drawn at time t. By updating the BMU and other units in the neighborhood, the distance between the BMU and the neighbors are brought closer together. The neighborhood function consists of two parts, one that define the form of the neighborhood and the other is the learning rate.

$$h_{ci}(t) = h(\|r_c - r_i\|, t) \propto (t) \qquad (2)$$

Where: r_c is the location of the winner unit; r_i is the location of the unit i on the grid map and $\propto (t)$ is the learning rate factor over minimum time t interval. At this stage the map converge to an inactive stage which approximates the probability density function of the high dimensional input data. The learning rate and the neighborhood proceed by time until convergence. Once the maps are trained, usually the concept of BMU which is used to facilitate the labeling of the consequent levels of fine tuning and refinement for the sake of tracing the related and diverse patterns.

The objective of the SOM visualization component is to render the SOM text file to a graphical representation. In SOM cluster files, the problem arose with neighboring neurons which are out of clusters and did not reflect exactly the severity of attack-ness in network connections [9]. That is because a network attack may not happen at a single action such that one massive attack may be start by seemingly innocuous or by small probe actions to take place [3]. In SOM classification process per example in [28], a genetic or clustering algorithm was used at certain attack zone to classify each attack by class whereas suspicious neurons which near the attack zone or out of the cluster area are not analyzed and remain suspicious were they might be benign. As one potential solution to this problem in the hierarchical SOM [2], they consider the potential of studying the domain knowledge of features to be applied to the whole SOM concepts.
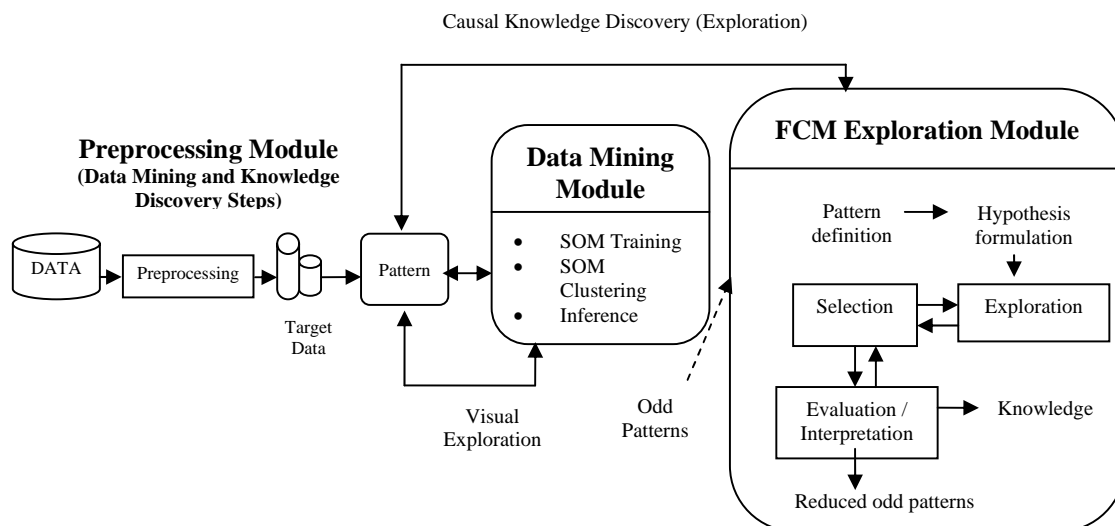
Figure 3. The classifiers anomaly detection system.

To increase the correlation among the neurons in the produced map grid, we minimize the neighborhood function and the learning rate by considering the minimum time interval according to the following rule:

In this study, we suggest an improvement to this process by considering the domain knowledge of particular neurons (odd neurons). Therefore, we used the FCM to calculate the severity/relevance of odd concepts (neurons) to attacks. Thus,

benign concepts can be dropped and/or others can be addressed as a potential risk of error caused.

C. FCM Exploration Module

The multi-classifier intrusion detection model is a defense-in-depth network based intrusion detection scheme. The model utilizes the domain knowledge of network data to analyze the packet information. Based on the analysis given, benign packets are dropped or blocked and high risk packets are to be highlighted or blocked using a causal knowledge reason in FCM.

The flowchart of the overall exploration steps are further illustrated in Figure 4. In this module the received data attributes will be carried out for fine-tuning in the FCM framework. As such, the neurons which represent low effect or less correlated to other attack like neurons are dropped or ignored and the high suspicious nodes are to be highlighted. Table I shows the degree of the effect and the value trace which represent the relations between neurons.

TABLE I.   EFFECT AND RELATION VALUE TRACE

Effect	Value Trace
Normal	0
Slight	0.2
Low	0.4
Somehow	0.6
Much	0.8
High	1

Initializing the FCM includes the definition of the FCM concepts and building the relations among these concepts by building a global matrix [4, 5]. However, in order to build that matrix we have defined the weight of odd neurons according to the total effect factor $Un(x)$ and the grade of causality W_{ij} between the nodes C_i and C_j according to the following assumptions:

1.   If $C_i \neq C_j$ and $E_{ij} \rangle 0$ then $W_{ij}^{+} = \max\{E_{ij}^{t}\}$
2.   If $C_i \neq C_j$ and $E_{ij} \langle 0$ then $W_{ij}^{-} = \max\{E_{ij}^{t}\}$
3.   If $C_i \neq C_j$ then $E_{ij} = 0$ and W_{ij} is zero

Each feature parameter of odd neurons is measured based on a comparison criteria to detect the interrelation between neurons i.e. determine the attack detection. To calculate the abnormality factor per packet we need to estimate the effect value of each feature parameter. The total degree of abnormality of odd neurons is calculated according to the total effect factor, the evaluation criteria illustrated in section IV.

The task of FCM is to determine the causal relationship between the suspicious or odd neurons noted by SOM to quantify causal inference process. By quantifying the causal inference process we can determine the attack detection and the severity of odd neurons such that neurons with low causal relations to be dropped or ignored. Using factors, rules and effect values we can estimate the total degree of effect value and hence the abnormality per packet. The effect parameters

and value trace are calculated according to [29]. The estimated the total degree of abnormality and attack detection per packet are considered according the following rule:

$$Un\ (x) = \sum_{i=1}^{n} E_i \qquad (3)$$

Where: $Un(x)$: Abnormality per packet

E_i : Effect value of packet

n : Total feature number of abnormality
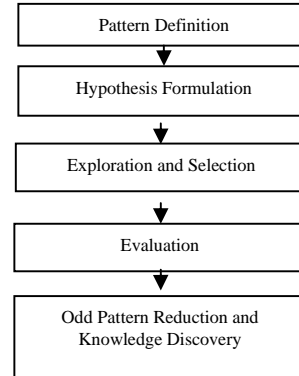
Figure 4.   The knowledge discovery module steps.

Once the abnormalities per un-clustered packets are calculated, the low malicious packets are dropped or ignored and the rest are considered as concepts in the FCM. It is now important to measure the effect/influence value among the suspicious concepts to determine the path of the existing or ongoing attack. If the effect value is zero then there is no relationship among these concepts. Figure 5 illustrates the FCM algorithm.

Algorithm	FCM
Inputs	SOM alerts, data patterns
Outputs	Reduced alerts, Replicate attack and normal connections
Method	
1	Define the number of odd neurons (SOM Alerts)
2	Define the number of concepts derived from data patterns
3	Call FCM initialization
4	Calculate the abnormality per neuron
5	Drop neurons if the abnormality is low
6	Until convergence show the link of related factors
End	

Figure 5.   FCM algorithm.

Once the normal and abnormal (attacks) data are replicated then it is time to explore the causal relationship of attack and alerts data to settle on the related factors and concepts. By doing so, we can further increase the detection accuracy and discover more related factors effectively. The knowledge discovery steps are illustrated in Figure 6.

1	Define attack and alerts patterns
2	Select atterns with probability P
3	Pick concepts with related similarity measures
4	Test correlations
5	Drop benign concepts (low correlation measures)
6	Highlight the rest as an attack or major error

Figure 6.   Exploration steps.

Table II gives a brief summary of the framework modules and its processes details.

TABLE II.   EXPLORAION MODULES SUMMARY

Module	Tasks	Method	Input	Output
Preprocessing	Data preprocessing	Standard and raw data selection	Data file	Cleansed and processed data file
Data Mining	Data mining	Cluster analysis	Data file	Report data files
Exploration	Pattern Discovery	Exploration	Data files, SOM files	Report data files

IV.   THE FEATURE SELECTION PROCESS

There are 41 features defined for every connection record. A complete listing of the sets of features defined for the connection records is given in the three tables below. The tables contain the list of 41 features available in the KDD'99 dataset [30]. These features are the intrusion detection dataset variables which are used for most of the IDS development and testing environment.

TABLE III.   BASIC FEATURES OF INDIVIDUAL TCP CONNECTIONS

# No	Feature name	Type
1	duration	continuous
2	protocol_type	discrete
3	service	discrete
4	src_bytes	continuous
5	dst_bytes	continuous
6	flag	discrete
7	land	discrete
8	wrong_fragment	continuous
9	urgent	continuous

TABLE IV.   CONNECTION FEATURES

# No	Feature name	Type
10	hot	continuous
11	num_failed_logins	continuous
12	logged_in	discrete
13	num_compromised	continuous
14	root_shell	discrete
15	su_attempted	discrete
16	num_root	continuous
17	num_file_creations	continuous
18	num_shells	continuous
19	num_access_files	continuous
20	num_outbound_cmds	continuous
21	is_hot_login	discrete
22	is_guest_login	discrete

TABLE V.   TIME-BASED FEATURES

# No	Feature name	Type
23	count	continuous
24	serror_rate	continuous
25	rerror_rate	continuous
26	same_srv_rate	continuous
27	diff_srv_rate	continuous
28	srv_count	continuous
29	srv_serror_rate	continuous
30	srv_rerror_rate	continuous
31	srv_diff_host_rate	continuous
32	Dst_host_count	continuous
33	Dst_host_srv_count	continuous
34	Dst_host_same_srv_rate	continuous
35	Dst_host_diff_srv_rate	continuous
36	Dst_host_same_src_port_rate	continuous
37	Dst_host_srv_diff_host_rate	continuous
38	Dst_host_serror_rate	continuous
39	Dst_host_srv_serror_rate	continuous
40	Dst_host_rerror_rate	continuous
41	Dst_host_srv_rerror_rate	continuous

A typical anomaly-based detection system works on the notion that abnormal behaviors and activities are different enough from normal (legitimate) behaviors profile. Once the normal and abnormal (attacks) data are replicated then it is time to explore the causal relationship of attack and alerts data to settle on the related factors and concepts. By doing so, we can further increase the detection accuracy and discover more related factors effectively. Our approach here is to study the probability of action of odd patterns and check for their correlation such that the higher correlation the higher related factors. The comparison criteria take the values between 0 and 1 such that:

If X and Y are two different concepts then:
$$Prob(X) = P(x) \tag{4}$$
$$Prob(Y) = P(y) * S \tag{5}$$

Where: S is the similarity ratio between X and Y. The concepts were defined according to the similarity ratio according to the following assumptions:

$$If\ X = Y\ then\ S = 1$$
$$If\ X <> Y\ then$$
$$S = 0$$

In other words, certain alerts can be assumed as an attack if they have similar probability of action. The similarity factor can be calculated according to the following rule:

$$S_f = \begin{cases} 1 & X = S_i \\ 0 & X \neq S_i \end{cases} \tag{6}$$

Where: S_i is a feature of set S and X is a comparison. The availability of features was estimated according to the following rule:

$$A_f = \begin{cases} 1 & X \in S \\ 0 & X \notin S \end{cases} \qquad (7)$$

Where: S is the set of features and X is a comparison. In this study, we have used the same testing data procedures as in (DARPA, 1998 and 1999) [30, 31]. We have extended that into evaluating the online generated (dumped) data patterns from CS USM computer forensic research lab as well as considering data features optimization and selection for a comprehensive evaluation as shown in the evaluation criteria diagram in Figure 7.
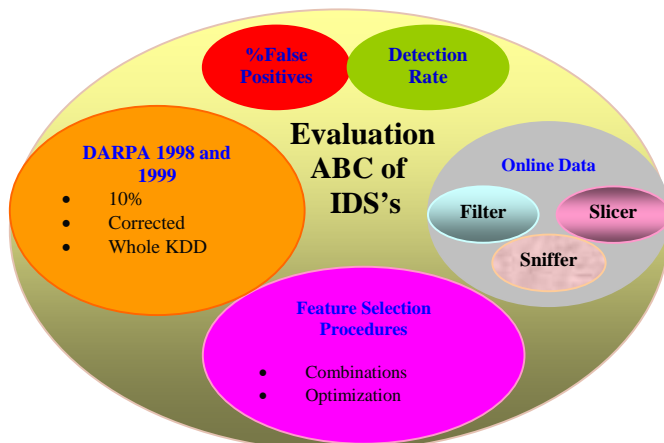

Figure 7.   IDS evaluation ABC.

In Figure 7, we have highlighted the most typical evaluation ABC for any IDS technology. It is also important to mention that there are many issues available regarding these evaluation ABC but we disregard to testing environment issues as the research mainly focus on the development of the inference engine components i.e. the multi-classifier system. Moreover, we have considered very important issue here which is the data feature selection procedure. The reason is due to the fact that most of the existing IDS technologies use all or none systematic patterns of data attributes to discover known or unknown intrusive activities; which will result in a lengthy intrusion detection process or even degrade the evaluation criteria of the IDS [26].

As stated earlier, the suggested approach is anomaly-based on causal knowledge reasoning. Using FCM, we attempts to diagnose and direct network traffic data based on relevancy to attacks and attack free connections. The SOM-FCM approach is a defense-in-depth intrusion detection model which utilizes the domain knowledge of network data to analyze packet information. The SOM and FCM schemes in combination and in isolation can be further modeled as an inference engine component for anomaly intrusion detection.

V. PERFORMANCE EVALUATION AND DISCUSSION

The proposed classifiers system architecture is two layers system architecture. The SOM layer trained to cluster the input data for each connection type. The layer of the SOM is belonging to one of the five classes of the dataset and will provide an output relevant to specific class. The second layer is the FCM framework. This layer uses a causal reasoning to measure the severity of odd neurons (SOM alerts). The main advantage here is to call attention to how domain knowledge of neurons (network packets) can contribute on tracing new attacks or find path of on-going or existing attacks.

In this study, we have used the most popular IDS evaluation data in which most of researchers aware of and used for evaluating their research, the KDD Cup 1999 intrusion detection contest data [30] followed by the success of the 1998 DARPA Intrusion Detection Evaluation Program by MIT Lincoln Labs (MIT Lincoln Laboratory) [31]. Table VI gives a sample distribution of the KDD'99 datasets.

TABLE VI.   SAMPLE DISTRIBUTION OF KDD'99 DATASET

Dataset Name	Normal	DOS	Probe	U2R	R2L	Total
Whole KDD	19.8	79.3	0.84	0.001	0.02	4,898,430
10% KDD	19.79	79.2	0.8	0.01	0.2	494,020
Corrected Test	19.58	73.9	1.3	0.02	5.2	311,029

The dataset contain 41 attributes for each connection record plus one class label and 24 attack types which fall into four main attack categories [32] as follows:

1. Probing: surveillance attack categories
2. DoS: denial of service
3. R2L: unauthorized access from a remote machine
4. U2R: unauthorized access to local super user (root) privileges

The dataset was established to evaluate the false alarm rate and the detection rate using the available set of known and unknown attacks embedded in the data set [33]. We have selected subsets from the so-called corrected.gz, 10% KDD, and the whole KDD files for testing purpose.

The selected subsets contain records with non zero values because some attacks are represented with few examples and the attack distribution in the large data set is unbalanced. Considering the whole data set degrades the IDS performance evaluation and result in boredom and lengthy detection process. However, the collection, preprocessing and calculation of false and true alert of test data are followed as in [3] according to the following assumptions:

FP: the total number of normal records that are classified anomalous
FN: the total number of anomalous records that are classified as normal
TN: the total number of normal records
TA: the total number of attack records
Detection Rate = [(TA-FN)/TA] * 100
False Alarm Rate = [FP/TN] * 100

In all cases, we run our experiment on a system with 2.667 GHz Pentium4 Processor 506 and 256MB PC3200 DDR400 RAM Running Windows XP.

For comprehensive evaluation, the performance measures of SOM, FCM and the combination of SOM-FCM are tested with specific and all features populations. As we have learned from the initial population testing, the dataset in 10% KDD simulates the entire or whole KDD. Therefore, we have conducted the test with only 10%KDD and the so-called CorrectedTest data.

Initially, we have considered the feature selection based on content, time and connection features for comprehensive experiments as shown in the following tables. First, we have tested the SOM and FCM in isolation and then the combined SOM-FCM approach. From the obtained results we have noticed that there is a prime importance for the connection features for detecting various attacks such that they can help in reduces false positives in most of the cases.

TABLE VII.   EXPERIMENTAL RESULTS FROM SOM MULTIPLE LAYERS USING CORRECTED TEST DATA

Feature Set	# Records	Normal Rate (%)	# Detection Records		
			FP rate (%)	FN rate (%)	Detection Rate (%)
Connection	311,029	19.58	15.7	0.002	99.3
Content	311,029	19.58	27.45	000	100
Time	311,029	19.58	41.5	0.002	99.3
All Features	311,029	19.58	20.0	0.012	93.8
Overall			26.16	0.004	98.10

TABLE VIII.   EXPERIMENTAL RESULTS FROM FCM USING CORRECTED TEST DATA

Feature Set	# Records	Normal Rate (%)	# Detection Records		
			FP rate (%)	FN rate (%)	Detection Rate (%)
Connection	311,029	19.58	9.1	000	100
Content	311,029	19.58	000	000	100
Time	311,029	19.58	1.0	000	100
All Features	311,029	19.58	11.34	000	95.36
Overall			5.36	000	98.84

TABLE IX.   EXPERIMENTAL RESULTS FROM SOM-FCM USING CORRECTED TEST DATA

Feature Set	# Records	Normal Rate (%)	# Detection Records		
			FP rate (%)	FN rate (%)	Detection Rate (%)
Connection	311,029	19.58	15.51	9..82	96.63
Content	311,029	19.58	000	4.68	60.14
Time	311,029	19.58	000	4.58	60.45
All Features	311,029	19.58	9.520	6.44	90.86709
Overall			6.25	3.92	77.02

TABLE X.   EXPERIMENTAL RESULTS FROM SOM-FCM USING 10% KDD DATA

Feature Set	# Records	Normal Rate (%)	# Detection Records		
			FP rate (%)	FN rate (%)	Detection Rate (%)
Connection	494,020	19.79	0.033	000	100
Content	494,020	19.79	0.002	000	100
Time	494,020	19.79	0.001	000	100
All Features	494,020	19.79	18.07	000	100
Overall			4.52	000	100

All of the above mentioned features and partitions of features illustrated in Section V have been applied to the inputs of the SOM, FCM, and SOM-FCM using single and multiple SOM classifiers. Initially, these methods are assumed to work as single class detectors for the input data patterns. As mentioned earlier, the SOM-FCM is a defense-in-depth anomaly intrusion detection scheme. For classification purpose, the classifiers parameters should be trained according to the nature of the given training data and then used to classify and categorize the data.

Since we employ the SOM classifier for each attack class category at the first layer of the SOM-FCM intrusion detection model, the number of SOM classifiers can be manipulated for testing purpose. Then the classifiers system can work as signature based classifiers system and each classifier work as anomaly detector. This flexibility gives the proposed soft computing components wide range of detection abilities and thorough understanding of the training data. According to [1], anomaly detectors perform better than misuse detectors over KDD'99 dataset using various machine learning algorithms. One explanation to this might be due to the complex distribution of the training samples and the embedded attack patterns in the KDD'99 data [34].

Due to this reason, we have randomly selected data partitions based on four attacks categories and normal data in order to test the proposed components specific to particular attacks categories. These selected data partitions are taken from the so-called 10% KDD dataset as it represent most of the attacks categories and emulates the whole KDD data. In addition, the 10% KDD had a very large number of records and hence requires long training time. The data partitions are randomly selected and their distribution samples are illustrated on the following tables.

TABLE XI.   SAMPLE DISTRIBUTION OF THE FIRST SELECTED DATA SET

		Normal	Probe	DoS	U2R	R2L
SOM-FCM	Normal	1000	300	600	050	500
SOM-FCM	Probe	1000	300	600	050	500
SOM-FCM	DoS	1000	300	600	050	500
SOM-FCM	U2R	1000	300	600	050	500
SOM-FCM	R2L	1000	300	600	050	500

TABLE XII.   SAMPLE DISTRIBUTION OF THE SECOND SELECTED DATA SET

		Normal	Probe	DoS	U2R	R2L
SOM-FCM	Normal	15000	3000	6000	100	5000
SOM-FCM	Probe	15000	3000	6000	100	5000

SOM-FCM	DoS	15000	3000	6000	100	5000
SOM-FCM	U2R	15000	3000	6000	100	5000
SOM-FCM	R2L	15000	3000	6000	100	5000

In Figure 8, we have demonstrated the classification and detection rates of the overall tests data cases which synopsis the overall tests results in the majority of cases. In fact, to obtain the desired detailed results about some cases we need to run tests over 10 times for particular case to show all possible graphics.
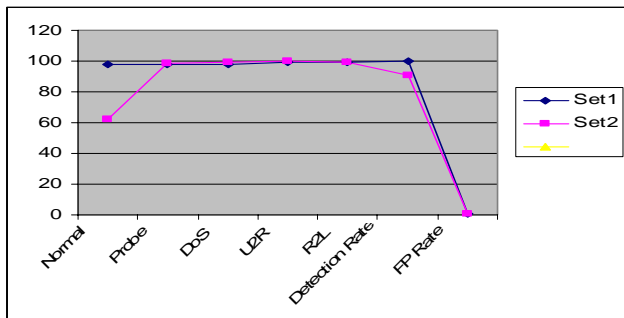

Figure 8.   SOM-FCM classification and detection rates.

From the obtained results, we can notice that SOM-FCM method can give high detection and classification rates in majority of cases. It is also clear that the higher the number of data samples the lower detection of normal data patterns. One explanation to this result could be the noise and the irregular patterns of attack and normal classes embedded in the dataset [35]. The evaluation presented in Figure 9 shows the detection records versus the false positive records in majority of cases. The figure shows the significantly decrease of false positives and the detection improvements using SOM-FCM.
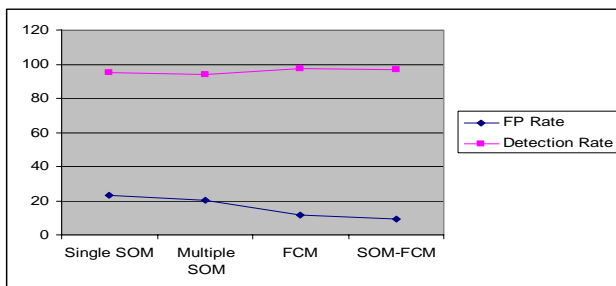

Figure 9.   SOM-FCM Detection rates vs. false positives rates.

The performance measure of SOM, FCM and the combination SOM-FCM with specific features populations was tested because of reasons mentioned in [36]. We considered the feature selection based on content, time and connection according to Minnesota IDS, MINDS [37]. It was clear that there is a prime importance of the connection features in various attack detection such that they can help in reduce false positives in most of the cases. We also notice that the SOM method triggers more false positives and false negatives without the basic features as compared with FCM and SOM-

FCM. As expected, there are exists a very complex relation among the data features. This condition goes worse when we detecting patterns from large dataset. Therefore, the test data must be always reduced for processing in intrusion detection. We believe that by reducing to the minimum number of feature we can significantly improve the classification, training time and hence improve the detection process. Figure 10 shows the overall assessment based on the specific features selection method used.
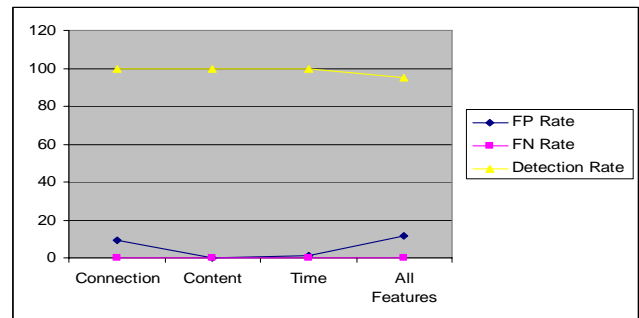

Figure 10. Detection records using various features in majority of cases.

We have conducted a test for the online generated data. The objective of this test is to show the effectiveness and the suitability of the approach for real-time intrusion detection. Typically, the sniffer components collect network traffic data in a promiscuous mode which make the whole network data has same factors of being suspicious or normal. This condition usually triggers full detection rate of all cases. However, normal traffic data present a few false positives due to some factors that are not representing the simultaneous connections such that features occurrences and relevancies from live network generated data are not balanced over a certain period of time. In this study, the content of the packet headers are used such as (TCP, UDP, and ICMP) and the features are portioned regard to the packet headers accordingly. For the real-time intrusion detection test, the collected network data are initially treated as normal while we running our antivirus software in parallel during the test which later realized as anomalous.

Figure 11 shows the overall online detection records. The figure shows that the detection rates were almost the same for most of the detection cases and the normal treated online generated data was detected as anomalous. One explanation to this situation might be the irregular relevance of the data patterns and the noise of the network traffic flow.
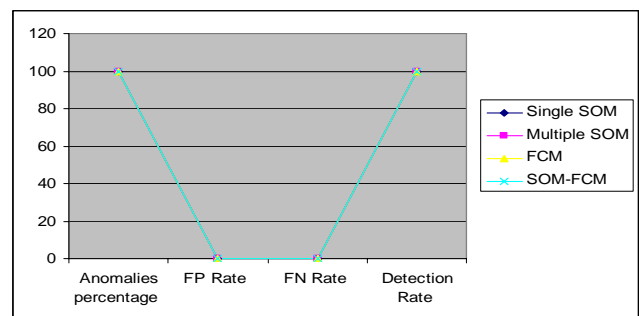

Figure 11. Online detection records.

The experimental results show that causal reasoning is a vital approach for intrusion detection. It was also clear that studying the probability of action of odd and attack neurons is of prime importance in order to trace attacks details and reduce number of false positives. We believe that further improvement on SOM-FCM structure will improve the detection accuracy and expose more information about attacks details. For future research, experiment should be done on comprehensive real-time environment and investigating methods for intelligent features selection and presentation.

VI. RELATED WORK

Most of the intrusion detection research focuses on the detection and classification methods. Unfortunately, existing systems failed in completely assessing the false alerts generation problem as well as the features and data attributes selection dominantly.

Recently, the interest on artificial intelligence (AI) techniques and data mining applications have received greater attention; particularly the use of unsupervised leaning methods as they have the ability to address some of the short comings for IDS [11]. This is also helps to achieve the ultimate goal for the IDS i.e. the capability of novelty detection. The unsupervised learning method so-called self organizing maps (SOM) has represented an excellent performance for sensors work on unsupervised learning mode [2] as well as it is efficient for real-time intrusion detection [12]. However, in order to refine the process and achieve better detection and performance, extra efforts are required. On the other hand, current trends on IDS didn't simply go for novelty detection but also to improve the reliability issues in term of detection, false positives, adaptability, speed and real-time issues.

To our best knowledge, existing studies on causal knowledge acquisition for intrusion detection are very limited. However, our work was also motivated by the work done recently on the intelligent IDS prototype [13] and the probe detection system (PDSuF) prototype [14]. The proposed intelligent IDS system [13] use fuzzy rule based and FCM as decision support tools and inference techniques. The proposed decision engine analyzes both misuse and anomaly modules information and combine both results for generating the final reports. For misuse information, the decision engine assesses the results from different misuse modules in order to capture misuse scenario. The anomaly detection module information is represented by neural networks as neurons, weights and relationship between the nodes.

The probe detection system (PDSuF) prototype [14] uses FCM for intrusion detection. In the proposed system, the decision module use FCM to capture and analyze packet information to detect SYN flooding attacks using a conventional FCM to measure the effect value based on the weight value between different variable events. Later, the decision module measures the degree of risk of DoS and trains the response module to deal with attacks. However, our approach is different from these approaches in such a manner that the suspicious events are generated from the flow of network packets depending on relevancy factors and causal relations among these factors using the FCM framework.

Based on the domain knowledge of network data, the proposed FCM framework uses a causal reason to measure the severity on network data. False positive alerts have been addressed by various studies for real-time operations [11, 12, and 15]. These studies whether are concentrating on speed detection or certain method improvement. On the other hand, false positive alerts have been addressed by various studies at sensor level [16, 17, and 18] by improving the sensor outputs. These studies whether are too general or concentrate on certain product improvement. Moreover, false alerts have been tackled at higher levels of the IDS operations. One such prototype is the Toolkit for Intrusion Alert Analysis [19], and the Intrusion Alert quality Framework [20] that uses certain quality parameters to improve the false positives using DARPA 2000 data set.

The various techniques used include data mining [21], AI techniques [13], fuzzy logic [22], neural networks [23] and neuro-fuzzy approach [24]. These techniques and approaches work on logs/alerts directly and indirectly by building new strategies to tackle intrusions of various types to improve the detection process. Other reported research work tackled the detection accuracy from the features selection perspective by ranking features subsets to represent different type of attacks [25, 26]. Here, we have considered features selection to be self extracted and learned.

In this study, we have established a link between SOM, FCM and used the combination for building better IDS sensor. Also the initial application of FCM and SOM architecture to the IDS problem has been reported [13, 2]. The focus of this research work will be on how practitioners can answer to specific elements or issues regarding the internal properties of events of SOM and FCM that have a specific influence on the performance of SOM-based intrusion detectors. The issues addressed in this study highlight question on how to eliminate ambiguities of odd neurons by extracting and presenting the most related features and factors.

The immediate result of this research is to improve the detection deficiency issue in the SOM-based IDS sensors by reducing the false alerts and increasing the detection accuracy at the sensor level. We believe that the biggest challenge here is to develop an intelligent inference engine to defense-in depth i.e. able to deal with uncertainty and detect novel attacks with low rate of false alerts. Moreover, any optimal solution of an adaptive IDS system should provide the means of real-time detection and response as well as high level trust among the IDS components.

REFERENCES

[1] A.N. Toosi, and M. Kahani, "A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers," Computer Communications 30(2007) 2201-2212.

[2] H. Gunes Kayacik, A.N. Zincir-Heywood, and M.I. Heywood, "A hierarchal SOM-based intrusion detection system," Engineering Applications of Artificial Intelligence 2006. doi:10.1016/j.engappai.2006.09.005

[3] S.T Sarasamma, Q.A. Zhu, and J. Huff, "Hierarchal kohonenen net for anomaly detection in network security," IEEE Transactions on Systems, Man, and

Cybernetics-Part B: Cybernetics, 35(2), 2005, pp. 302-312.

[4] C.D. Stylios, and P.P Groumpos, "Mathematical formulation of fuzzy cognitive maps," Proceedings of the 7th Mediterranean Conference on Control and Automation (MED99), Haifa, Israel, 1999.

[5] B. Kosko, "Fuzzy cognitive maps," International Journal of Man-Machine Studies, Vol. 24, 1986, pp. 65-75.

[6] T. Kohonen, Self-organizing maps. Third ed. Springer, Berlin, 2000.

[7] T.S. Sobh, "Wired and wireless intrusion detection system: Classification, good characteristics and state-of-art," Computer Standards & Interfaces 28 (2006) 670-694.

[8] R.G. Bace, Intrusion Detection. ISBN 1-57870-185-6, 2001.

[9] A. Seleznyov, and Puuronens, "Anomaly inrtrusion detection systems: Handling temporal relations between events," Proceeding of the 2nd international workshop on recent advances in intrusion detection (RAID'99).

[10] M. Jazzar, and J. Aman, "Using fuzzy cognitive maps to reduce false alerts in SOM-based intrusion detection sensors," 2nd Asia International Conference on Modeling and Simulation (AMS2008), pp. 1054-1060.

[11] M. Amini, R. Jalili, and H.R Shahriari, "RT-UNNID: A practical solution to real-time network-based intrusion detection using unsupervised neural networks," Computers & Security 25(2006) 459-468.

[12] W. Wang, X. Guan, X. Zhang, and L. Yang, " Profiling program behavior for anomaly intrusion detection based on the transition and frequency property of computer audit data," Computers & Security 25 (2006) 593-550.

[13] A. Siraj, R.B. Vaughn, and S.M. Bridges, "Intrusion sensor data fusion in an intelligent intrusion detection system architecture." Proceeding of the 37th Hawaii International Conference on System Sciences, 2004.

[14] S.Y. Lee, Y.S. Kim, B.H. Lee, S. Kang, and C.H. Youn, "A probe detection model using the analysis of the fuzzy cognitive maps," International Conference on Computational Science and its Applications, ICCSA (1) 2005, pp. 320-328.

[15] W. Jiang, H. Song, and Y. Dia, "Real-time intrusion detection for high-speed networks," Computers & Security 24 (2004) 287-294.

[16] K. Timm, "Strategies to Reduce False Positives and False Negatives in NIDS," SecurityFocus Article, 2001. Available: www.securityfocus.com/infocus/1463

[17] M.J. Ranum, "False Positives: A User's Guide to Making Sense of IDS Alarms," ICSA Labs IDSC, 2003.

[18] M. Norton, and D. Roelker, "Snort 2.0 Rule Optimizer," Sourcefire Network Security White Paper, April 2004.

[19] TIAA, A Toolkit for Intrusion Alert Analysis (Version 0.4). Available: http://discovery.csc.ncsu.edu/software/correlator/ver0.4

[20] N.A. Bakar, B. Belaton, and A. Samsudin, "False Positive Reduction via Intrusion Alert Quality Framework," 13th IEEE international Conference on Networks, Kuala Lampur, Malaysia, Vol. 1, 2005, pp. 547-552.

[21] W. Lee, S.J. Stolfo, and K.M. Mok, "Adaptive intrusion detection: A data mining approach," Artificial Intelligence Review 14(6), 2000, pp. 533- 567.

[22] J.E. Dickerson, J. Juslin, O. Koukousoula, and J.A Dickerson, "Fuzzy intrusion detection," IFSA World Congress and 20th North American Fuzzy Information Processing Society (NAFIPS) International Conference, Vancouver, British Columbia, 2001.

[23] Y. Liu, D. Tian, and A. Wang, "ANNIDS: Intrusion detection system based on artificial neural network,"

Proceedings of the second international conference on machine learning and cybernetics, Xi'an, 2003.

[24] R. Alshammari, S. Sonamthiang, M. Teimouri, and D. Riordan," Using neuro-fuzzy approach to reduce false positive alerts," Fifth Annual Conference on Communication Networks and Services Research (CNSR'07), IEEE Computer Society Press, 2007, pp. 345-349.

[25] A.H. Sung, S. Mukkamala, "The feature selection and intrusion detection problems," In: M.J Maher (ed.) ASIAN 2004. LNCS, vol. 3321, pp. 468-482. Springer, Heidelberg (2004).

[26] A. Zainal, M.A. Maarof, and S.M. Shamsuddin, "Feature selection using rough-dpso in anomaly detection," In: O. Gervasi, and M. Gavrilova (eds.): ICCSA 2007. LNCS 4705, part I, pp. 512-524. Springer-Verlag Berlin Heidelberg 2007.

[27] O. Depren, M. Topallar, E. Anarim, and M.K. Ciliz, "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks," Expert Systems with Applications 29 (2005) 713-722.

[28] L. DeLooze, "Attack characterization and intrusion detection using an ensemble of self-organizing maps," Proceeding of the 2006 IEEE Workshop on Information Assurance, United States Military Academy, West Point, NY, 2006.

[29] M. Jazzar, and J. Aman, "An Approach for anomaly intrusion detection based on causal knowledge-driven diagnosis and direction," In: R. Lee (Ed.): Soft. Eng., Arti. Intel., Net. & Para./Distri. Comp., Studies in Computational Intelligence (ISC), Vol. 149, pp. 39-48. Springer Berlin / Heidelberg Press, ISBN: 978-3-540-70559-8.

[30] KDD Cup 1999 Data. Knowledge Discovery in Databases DARPA Archive. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[31] MIT Lincoln Lab, DARPA Intrusion Detection Evaluation Plan. Available: http://www.ll.mit.edu/IST/ideval/data/2000/2000_data_index.html.

[32] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," Journal of Network and Computer Applications, 2005.

[33] K. Kendall, "A database of computer attacks for the evaluation of intrusion detection systems," Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA, 1999.

[34] D. Song, M.I Heywood, and A.N Zincir-Heywood, "Training genetic programming on half a million patterns: An example from anomaly detection. IEEE Transactions on Evolutionary Computation 9 (3), pp. 225-239. doi: 10.1109/TEVC.2004.841683

[35] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozjur, and J. Srivatava, "A comparative study of anomaly detection schemes in network intrusion detection," Proceeding of the SAIM International Conference on Data Mining. San Francisco, CA.

[36] A. Abraham, and R. Jain, "Soft computing models for network intrusion detction systems," Studies in computational intelligence (SCI) 4, 191-207 (2005).

[37] L. Ertoz, E. Eilertson, A. Lazarevic, P. Tan, J. Srivastava , V. Kumar, P. Dokas, "The MINDS - Minnesota Intrusion Detection System," Next Generation Data Mining, MIT Press, 2004.

AccSearch: A Specialized Search Engine for Traffic Analysis

K. Renganathan

Computer Science and Engineering Department
SRM University
India
ranganath73@gmail.com

B. Amutha

Computer Science and Engineering Department
SRM University
India
bamutha62@gmail.com

Abstract— **AccSearch is a specialized web search engine to provide information about road accidents within Chennai, India and assist the traffic authorities, police, NGOs, lawyers, students and statistical bureaus. The people who are in need of road accident information for various reasons are very much struggling to collect the correct information under a single search. Special purpose search engines are designed to work on a particular domain which fill the gap where an all purpose search engine lacks. As the existing search engines cannot do the traffic search alone well for several reasons, we have designed a search algorithm using Markov chain, to provide the search information in a faster manner. The mathematical proof of our modified Markov chain algorithm shows that the speed and efficiency seems to be better in comparison with the existing search algorithms. As Markov chain can be used for prediction purposes, our search engine concentrates on one particular domain which is traffic analysis it will result in exact responses to the user queries and will lead to a greater amount of user satisfaction.**

Keywords; AccSearch; road traffic; accident;Markov chain;accident prediction;

I. INTRODUCTION

Road accidents are the major problem in many countries. It is a very series problem in the highways of India. Internet is grown very large. As it is very large and the information is scattered all around the world, search engines are the only medium through which the information can be accessed. But the relevancy of the search result is the major problem in search engines. Though popular search engines like Google perform well through their quality of page ranking algorithms still it is true that many questions remains unanswered up to their expected relevancy. Special purpose search engines are those search engines which attempt to answer those questions which are not answered or cannot be answered by an all purpose search engines.

This project is an effort to create a comprehensive special purpose search engine which will support with accurate responses with maximum possible relevancy till the very last URL result for any queries pertaining to the road accident details within the Chennai city. It is aimed to provide a high dependency to the user. It covers the entire accident data which occurred in the four National Highways around Chennai. It will provide information about the accident occurred in the day and night around the highways. This information is used to do the historical collection of data. This traffic search engine can be later connected to the all purpose search engines to add up the searching power and efficiency.

Markov chain algorithm is used to improve the performance and speed up. Markov chains are well known for the performance tuning and prediction.

Adding the information with the available information on the internet is the fruit of this work. By providing some more information with the already available information some sectors will be highly benefited. Those include Police, NGOs, Statistical Bureaus and Universities to name a few. It will provide a greater benefit to the society.

A.   Literature Survey / Related Works

Sergey Brin and Lawrence Page, *"The Anatomy of Large-Scale Hyper textual Web Search Engine" addressed the issue of developing a large scale search engine such as google but failed to address the issue of specialized search[1]*Sunny Lam, *"The Overview of Web Search Engines,"* addressed the issue of how the search engines find information in the Web and how they rank the pages according to the given query. It helps people perform Web searching easily and effectively. But it not address the issue of not getting the required information even after search[2].

Robert Steele, *"Techniques for Specialized Search Engines"* addresses the issue of the need for specialized search engine.[3]Z Xiang, K. Wober, DR. Fesenmaier, *"Representation of the Online Tourism Domain in Search Engines,"* Addresses the issue of increasing the search results in tourism domain using techniques.But failed to provide the lack of important information related to tourism on the web[4]. Z Xiang, Bing Pan, K. Wober, DR. Fesenmaier, *"Developing SMART- Search : A Search Engine to Support the Long Tail in*

Destination Marketing," address the issue of effective organizing over the internet to support travel information search. It didn't address the issue of how to increase the availability of travel information[5].

Karl W Wober, *"Domain Specific Search Engines,"* addresses the techniques involved in domain specific search. But doesn't address the issue of how to implement the domain specific search engine[6].YAN Hongfei, LI Jingjing, ZHU Jiaji, PENG Bo, *"Tianwang Search Engine at TREC 2005: Terabyte Track,"* address the issue of large amount of data transfer. It does not address the issue of improving the search results[7]. Jermy Ginsberg, Mathew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant *"Detecting influenza epidemics using search engine query data,"* address the issue of detecting influenza using the query data.It does not discuss about how avoid such epidemics using that data[8].Gang Luo, Chunqiang, Hao Yang, Xing Wei, *"MedSearch : A Specialized Search Engine for Medical Information,"* addresses the issue of how to help layman in medical search but not addresses the issue of relevancy among the medical information results[9].

Jianhan Zhu, Jun Hong, and John G. Hughes, *"Using Markov Chains for Link Prediction in Adaptive Web Sites,"* addresses the navigation problems in adaptive web sites. But it does not address link prediction from the past state to future state[10]. Junghoo Cho, Hector Gracia Molina, Lawrence Page, *"Efficient Crawling through URL Ordering,"* addresses the issue of in what order the crawler should visit URLs. But not addressed the issue of taking care of the missed URLs which are not came in order of the crawler[11].

Junghoo Cho, Hector Gracia Molina, *"The evolution of the Web and Implications of an Incremental Crawler,"*addresses the issue of incrementally updating the index. But not addressed the issue of updating the indexes randomly [12]Junghoo Cho, Hector Gracia Molina, *"Parallel Crawlers,"* address the issue of managing the indexing of ever growing web. It doesn't give the complete guidelines to construct parallel crawlers[13].Sanjay Kumar Singh, Ashish Misra, *"Road Accident Analysis : A Case Study of Patna City,"* addresses the issue of Road accidents in Patna city. But not addressed any road safety measures[14].G D Jacobs, Amy Aeron Thomas, *"A Review of Global Road Accident Fatalities,"* addresses the issue of deaths and injuries during accidents. But not addressed how public and private sector can act to prevent these injuries[15].

Ramasamy. N, *"Accident Analysis of Chennai City,"* addressed the issue of accident analysis of Chennai city. But not addressed how to avoid such accidents in future[16].Dinesh Mohan, *"Social Cost of Road Traffic Crashes in India,"* addressed the issue of cost of injuries and deaths. But not addresses how to eliminate those unwanted cost[17].P.Pramada VALLI, *"Road Accident Models for Large Metropolitan Cities of India,"* addressed the issue of preventing accidents by road accident model. But not addresses how to avoid accidents even

after the model has been built[18].Pachaivannan Partheeban, Elangovan Arunbabu, Ranganathan Rani Hemamalini, *"Road Accident Cost Prediction Model Using Systems Dynamics Approach,"* addressed the issue of reducing the cost of accident using developing model using systems dynamic approach. But not addresses that will it really lead to accurate cost prediction[19].

I. NEED FOR A SPECIALIZED SEARCH ENGINE

All purpose search engines are very broad and deemed to cover almost all domains in the world. Though this quality is an advantage it includes some inabilities too. The main factors which influence any search engine and create the specialized need are found and listed as below:

- Specialization
- Availability of Information
- Responsibility
- Time elapsed

A. Specialization

Though all purpose search engines support specialization of information in response to the user queries, but they are mainly meant for generalization of information. Curious search engines use the user queries which are unanswered or not properly answered with expected relevancy to enhance their system to answer well in feature. But at that point of time when user expects the right answer to his specialized queries he won't be able to get.

B. Availability of Information

All purpose search engines gather information from all around the web. It has tons of information to serve the users. It will answer the maximum of the user queries. But it won't be able to answer all the queries. Because it doesn't possess the information by its own. These search engines will struggle in answering queries which requests in depth details within a particular domain.

C. Responsibility

All purpose search engine tries through all the means to respond well for the user query and as well as update its information repository well to keep it fit for this activity. But it bears no responsibility to answer the queries positively. Hence, it is not sure for the user that his queries will be answered. It will be a trial and error process for him. All purpose search engines works with probability not with accuracy in this aspect. Some search engine may handle some searches with most probably high relevancy and for some other searches with less probability. This makes the user difficult to rely on such kind of search engines.

D. Time elapsed

Time elapsed in searching is the major factor which affect the interest of the user. When the time elapsed is more, it will

create a greater amount of dissatisfaction in users. It has been found that the users are spending hours or sometime days in searching some essential information among the web. After getting dissatisfied by their prolonged search they use to try some other means to get that information i.e., making a series of phone calls, trying in yellow pages, physically going to the concerned place to get that information etc. Hence, the efforts made to reduce this time elapsed will bring a giant leap in the development of the advanced search engines.

The specialized search engine is aimed to address the above factors which are not addressed by the all purpose search engines. Firstly it will concentrate on one domain and will have sufficient collection of information to answer all sorts of queries in that particular domain. As it is assured to answer all the queries within that domain the user can fully rely on it. Hence, it creates the full dependability to the user. The specialized search engine AccSearch will contain all needed information local to its domain. It will ensure the availability of all the essential information. It bears the responsibility for the information availability. It makes the user queries will be answered with full relevancy. It reduces the time elapsed in searching by the user. It will answer the very first query itself with full relevancy(whereas normally it needs many queries to obtain an information in an all purpose search engine). At maximum level he may need to try with very few queries. Finally he can finish off his search in few minutes instead of long time hassles. It has been found that there are regular users to search engine and they need to search for information for their day to day activities. We identified the target users for AccSearch. They are Police, NGOs, Statistical Bureaus, Lawyers, Students to name a few. There will be bundle of global users too. Once it attained perfection on its domain it will be made to crawl the whole www so that it will work specialized on its domain and generalized on all-purpose search

II. MODIFIED MARKOV CHAIN ALGORITHM FOR ACCSEARCH

A. Assumptions

- Types of vehicles : VT_1, VT_2, VT_3, VT_4, VT_5, VT_6, VT_7
- Types of accidents: F_1, F_2, F_3, F_4
- Time of accidents: tn_1, tn_2, tp_1, tp_2
- Number of accidents : N_i
- Search engine : S_1, S_0

Types of vehicles:

- VT_1 – Government Bus
- VT_2 – Private Bus
- VT_3 – Truck/Lorry
- VT_4 – Car/Jeep/Taxi/Tempo

- VT_5 – Two wheelers
- VT_6 – Three wheelers
- VT_7 – Others [bye cycle, bullock cart etc.,]

Type of accidents:

- F_1 – Fatal (Death)
- F_2 – Grievous Injury
- F_3 – Minor Injury
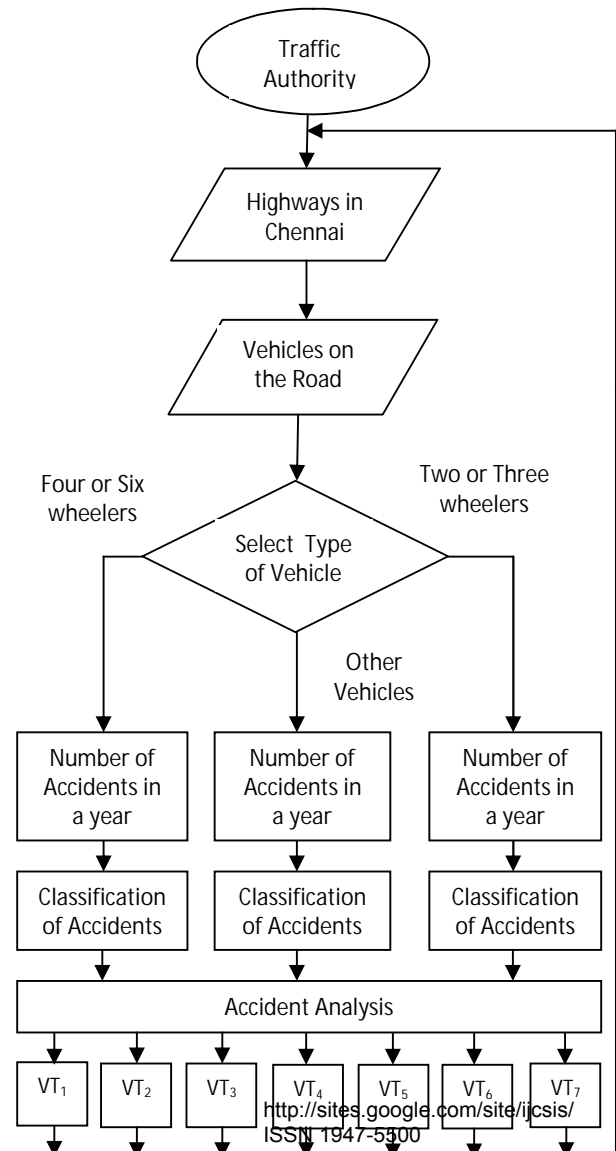- F_4 – Non Injury

Time of accidents:

Peak hours:

tp_1 – 8:30 AM to 9:30 AM
tp_2 – 5 PM to 6:30 PM

Normal hours:

tn_1 – 10 AM to 5 PM
tn_2 – 7 PM to 8 AM (Cargo)

B. The Process Flow of Accident Analysis

The equations which are the result of the accident analysis are given below. Each element in an equation represents the percentage of the type of accident occurred with respect to the total number of accidents by a particular vehicle type.

$VT_1 \rightarrow 21.69F_1 + 0.39F_2 + 51.67F_3 + 26.23F_4$

$VT_2 \rightarrow 16.43F_1 + 0.68F_2 + 48.63F_3 + 34.24F_4$

$VT_3 \rightarrow 19.74F_1 + 0.31F_2 + 42.63F_3 + 37.3F_4$

$VT_4 \rightarrow 7.27F_1 + 1.07F_2 + 58.37F_3 + 33.27F_4$

$VT_5 \rightarrow 11.31F_1 + 1.38F_2 + 81.14F_3 + 6.15F_4$

$VT_6 \rightarrow 7.92F_1 + 0.93F_2 + 80.18F_3 + 10.95F_4$

$VT_7 \rightarrow 32.46F_1 + 1.15F_2 + 47.53F_3 + 18.84F_4$

As the vehicle types VT_1 and VT_2 are similar types (Bus) and the VT_2 is available in only negligible amount and its effect on the accidents is very low these two types can be merged.

$VT_1 \& VT2 \rightarrow 19.6F_1 + 0.535F_2 + 50.15F_3 + 30.235F_4$

C. Algorithm

Now the algorithm may be expressed as follows:

If
  S_1.Vehicle = VT_1 & VT_2
        &
  S_1.Time = tp_1 & tp_2

  Type.Accident = S_0.[$19.6F_1 + 0.535F_2 + 50.15F_3 + 30.235F_4$]

Else If

  S_1.Vehicle = VT_3
        &
  S_1.Time = tn_2

  Type.Accident = S_0.[$19.74F_1 + 0.31F_2 + 42.63F_3 + 37.3F_4$]

Else If

  S_1.Vehicle = VT_4
        &
  S_1.Time = tp_1 & tp_2

  Type.Accident = S_0.[$7.27F_1 + 1.07F_2 + 58.37F_3 + 33.27F_4$]

Else If

  S_1.Vehicle = VT_5
        &
  S_1.Time = tp_1 & tp_2

  Type.Accident = S_0.[$11.31F_1 + 1.38F_2 + 81.14F_3 + 6.15F_4$]

Else If

  S_1.Vehicle = VT_6
        &
  S_1.Time = tp_1 & tp_2

  Type.Accident = S_0.[$7.92F_1 + 0.93F_2 + 80.18F_3 + 10.95F_4$]

Else If

  S_1.Vehicle = VT_7
        &
  S_1.Time = tp_1 & tp_2

  Type.Accident = S_0.[$32.46F_1 + 1.15F_2 + 47.53F_3 + 18.84F_4$]

End If

III. MATHEMATICAL MODEL OF THE ALGORITHM

The transition matrix has been constructed using these available results.

$V_1 \rightarrow VT_1 \& VT_2 \rightarrow 19.6F_1 + 0.535F_2 + 50.15F_3 + 30.235F_4$

$V_2 \rightarrow VT_3 \rightarrow 19.74F_1 + 0.31F_2 + 42.63F_3 + 37.3F_4$

$V_3 \rightarrow VT_4 \rightarrow 7.27F_1 + 1.07F_2 + 58.37F_3 + 33.27F_4$

$V_4 \rightarrow VT_5 \& VT_6 \& VT_7 \rightarrow 17.23F_1 + 1.153F_2 + 69.617F_3 + 11.98F_4$

A. Transition Matrix

$$
T = \begin{array}{c} \\ F_1 \\ F_2 \\ F_3 \\ F_4 \end{array}
\begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \\
\left[\begin{array}{cccc}
.196 & .1974 & .0727 & .1727 \\
.00535 & .0031 & .0107 & .01153 \\
.5015 & .4263 & .5837 & .69617 \\
.30235 & .373 & .3327 & .1198
\end{array}\right]
\end{array}
$$

The above traffic prediction analysis expressed in terms of transition matrix shows that the row represents fatality factors in correspondence with the vehicle classification.

In the same manner the columns of the matrix shows that according to the vehicle types the percentage of fatality occurred. As this is a real time classification which have been made in the Chennai city in the year 2008.

IV. Experimental Validation and Results

Predictions for V_1

Predictions for V_3

Predictions for V_2
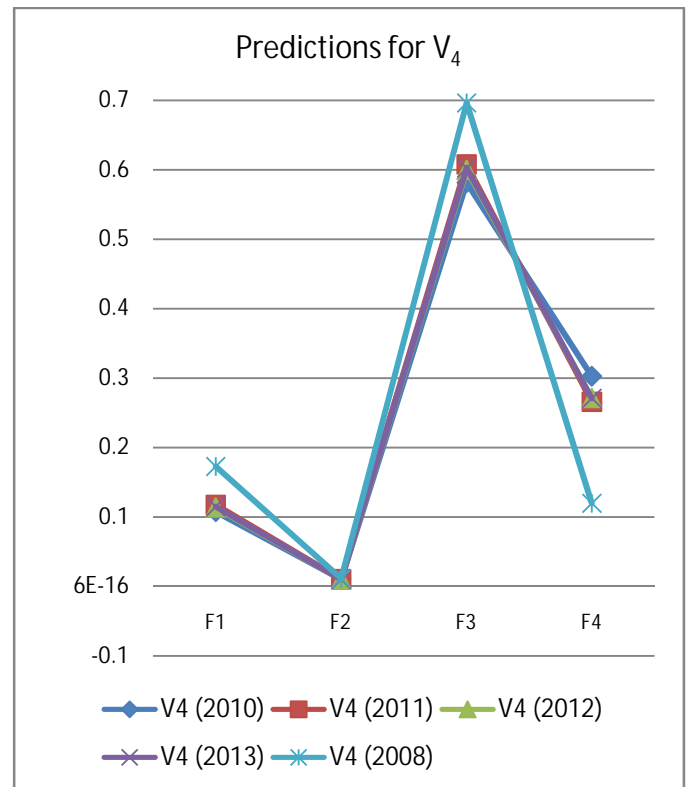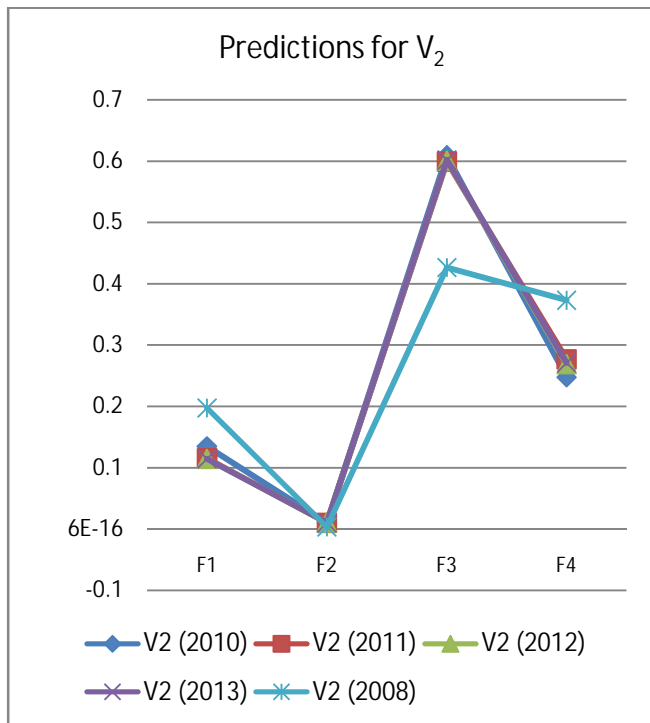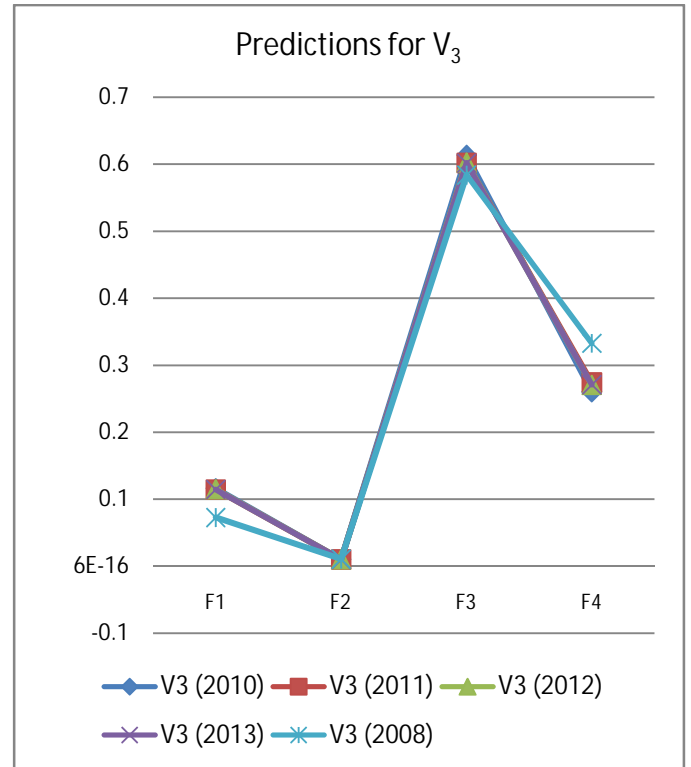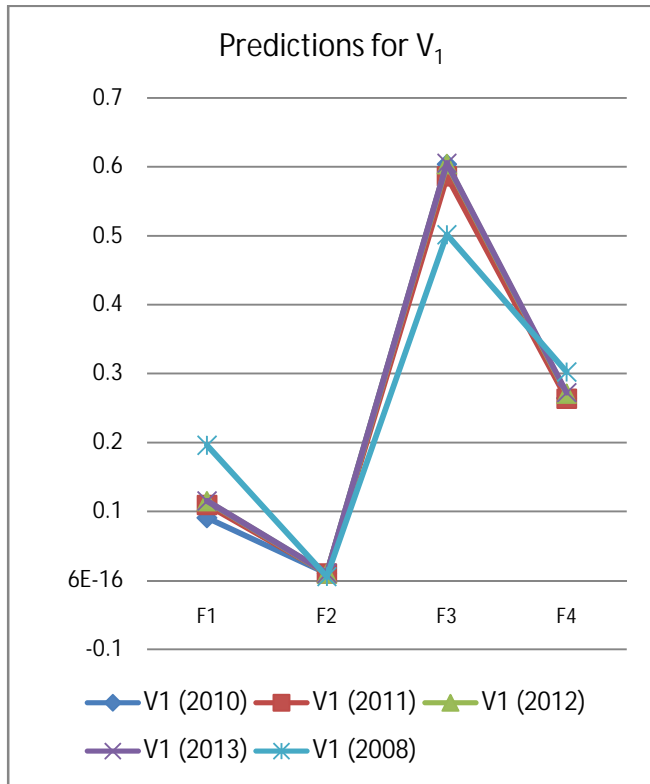
Predictions for V_4

Predictions for V_1

Predictions of F_1 of 2010 – 2013are less when compared to 2008 whereas 2008 shows a value of 19.6% but 2010-2013 are found as 9.07%, 10.93%, 11.49% and 11.52% respectively. In predictions of F_2 2008 shows a lower value that is 0.53% whereas 2010-2013 shows higher value i.e, 0.99%,1%, 1.02%, and 1.03% respectively. In the predictions of F_3 2010, 2012 and 2013 shows the higher values and the remaining shows the lower value. The values for 2010-2013 are 60.38%, 60.37 and 60.51 respectively. Similarly for 2011 and 2008 the values are 58.61% and 50.15%. Predictions of F4 states that the value of 2008 is somewhat raised and 2010-2013 are somewhat lowered. The values are 26.43%, 26.37%, 27.11% and 27.24% respectively whereas the 2008 value is 30.23%.

Predictions for V_2

Predictions of F_1 of 2011-2013 are showing nearby values where 11.53%, 11.45%, and 11.45% respectively whereas the values of 2010 and 2008 are different those are, 13.47% and 19.74% respectively. Predictions of F_2 shows that the values of 2010 -2013are almost similar, those are 0.99%, 1.01%, 1.02%, and 1.02%. But 2008 shows 0.31%. Predictions of F_3 shows that 2010 , 2012 and 2013 shows almost similar values those are 60.88%, 60.88% and 60.16%. The values of 2011and 2008 showing distinct such as 59.94% and 42.63%. Predictions of F_4 shows that 2010-2013shows almost similar values those are 24.74%, 27.66%, 26.9%, and 27.1%.

Predictions for V_3

Predictions of F_1states that 2010-2013 have almost similar values that is 11.63%, 11.44%, 11.5%, and 11.46% respectively whereas 2008 represents 7.27%. Predictions of F_2 is showing that 2011-2013 are same and 2010 is almost same that is 1.02 for 2011-2013 , 1.05 for 2010 and 1.07 for 2013. Predictions of F_3 states that values of 2010-2013 are almost same those are, 61.33%, 60.18%, 60.25%, and 60.16%. For 2008 it is 58.37%. Predictions of F_4 states that the values of 2010-2013 are almost same those are 26%, 27.43%, 27.07% and 27.09% respectively. But the value of 2008 bears 33.27%.

Predictions of V_4

Predictions of F_1 states that the values of 2011-2013 are almost similar values say 11.75%, 11.39%, and 11.49%. The year 2010 which is 10.74%. 2008 shows a value 17.27%. Predictions of F_2 states that the values of 2012 and 2013 are similar and 2011 is almost similar, 1.02%, 1.02% and 1.03%. 2010 shows 9.8% and 2008 shows 1.153%. Predictions of F_3 states that 2011-2013 and 2008 have almost similar values these have 60.79%, 59.99%, 60.25%, and 69.617% whereas 2010 shows a lower value 58.13%. Predictions of F_4 shows

that 2011-2013 are nearby values and 2008 and 2010 are distinct values. The values of 2011-2013 are 26.58%, 27.18%, and 27.09% respectively. The values of 2010 and 2008 are 30.25% and 11.98%.

V. CONCLUSION AND FUTURE WORKS

This paper presents AccSearch, a specialized web search engine for road accident information retrieval. It will aid the user group consisting of police, NGOs, statistical bureaus, lawyers, students and others who may require road accident information for their day to day activities. AccSearch is designed to be a scalable search engine. The primary goal is to provide a very high relevancy in search results.

In future this search engine will be enhanced as a semantic search engine by creating ontology for this domain.

REFERENCES

[1]. Sergey Brin and Lawrence Page, *"The Anatomy of Large-Scale Hypertextual Web Search Engine"*, Computer Networks 30(1-7): 107-117, 1998.

[2]. Sunny Lam, *"The Overview of Web Search Engines,"* O. Waterloo – University of Waterloo, 2001.

[3]. Robert Steele, *"Techniques for Specialized Search Engines,"* Proceedings of Internet Computing, 2001.

[4]. Z Xiang, K. Wober, DR. Fesenmaier, *"Representation of the Online TourismDomain in Search Engines,"* 47(2) 137 Journal of Travel Research, 2008

[5]. Z Xiang, Bing Pan, K. Wober, DR. Fesenmaier, *"Developing SMART- Search : A Search Engine to Support the Long Tail in Destination Marketing,"* www.ota.cofc.edu.

[6]. Karl W Wober, *"Domain Specific Search Engines,"* Wallingford, UK : CABI , 2006, www.tourism.wu.wien.ac.at.

[7]. YAN Hongfei, LI Jingjing, ZHU Jiaji, PENG Bo, *"Tianwang Search Engine at TREC 2005: Terabyte Track,"* Network and Distribution System Laboratory, School of Electronic Engineering and Computer Science, Peking University Beijing, China.

[8]. Jermy Ginsberg, Mathew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant *"Detecting influenza epidemics using search engine query data,"* Nature, vol. 457, 19 February 2009.

[9]. Gang Luo, Chunqiang, Hao Yang, Xing Wei, "*MedSearch : A Specialized Search Engine for Medical Information,*" IBM Research Journal, RC24205, March 6 2007.

[10]. Jianhan Zhu, Jun Hong, and John G. Hughes, "*Using Markov Chains for Link Prediction in Adaptive Web Sites,*" Soft-Ware 2002, LNCS 2311, pp 60 – 73, Springer-Verlog 2002.

[11]. Junghoo Cho, Hector Gracia Molina, Lawrence Page, "*Efficient Crawling through URL Ordering,*"

[12]. Proceedings of Seventh International Web Conference (WWW 98), 1998.

[13]. Junghoo Cho, Hector Gracia Molina, "*The evolution of the Web and Implications of an Incremental Crawler,*" Department of Computer Science, Stanford University, Stanford, CA 94305, USA.

[14]. Junghoo Cho, Hector Gracia Molina, "*Parallel Crawlers,*" Department of Computer Science, Stanford University, Stanford, CA 94305, USA.

[15]. Sanjay Kumar Singh, Ashish Misra, "*Road Accident Analysis : A Case Study of Patna City,*" Urbon Transport Journal 2(2): 60-85, 2001.

[16]. G D Jacobs, Amy Aeron Thomas, "*A Review of Global Road Accident Fatalities,*" RoSPA Road Safety Congress, Plymouth, UK , 3-7 March 2000, at http://www.transport-links.org.

[17]. Ramasamy. N, "*Accident Analysis of Chennai City,*" Working Paper 3, Centre for Road Safety, Central Institute for Road Safety, Pune 2001.

[18]. Dinesh Mohan, "*Social Cost of Road Traffic Crashes in India,*" Proceedings : First Safe Community Conference on Cost of Injury, pp 33-38, Viborg, Denmark, October 2002.

[19]. P.Pramada VALLI, "*Road Accident Models for Large Metropolitan Cities of India,*" IATSS Research, Vol.29, No.1, 2005.

[20]. Pachaivannan Partheeban, Elangovan Arunbabu, Ranganathan Rani Hemamalini, "*Road Accident Cost Prediction Model Using Systems Dynamics Approach,*" ISSN 1648-4142 print / ISSN 1648 – 3480 online 2008, at www.transport.vgtu.it

[21]. "*Road Safety Policy,*" Road Safety Book, Home, Prohibition and Excise Department, Government of Tamil Nadu April 2007.

[22]. Leslie Hogben, *Matrices,Digraphs,Markov Chains & Their Use by Google* , Bay Area Mathematical Adventures, Iowa State University and American Institute of Mathematics, February 27, 2008.

ANNEXURE:

1. Among the total number of accidents fatal accidents are low in numbers and the non-fatal are high in numbers. A sample trend is shown below:

Fatal is shown as (2005: 584) (2006 : 627) (2007 : 583) (2008:612)(2009:609)
Non fatal is shown as (2005:4427) (2006 :4657) (2007: 4277) (2008:5774)(2009:4575)

Where the total number of accidents of these categories are: 5177

2. Among the various vehicle types two wheelers is the more prone to accidents of fatal category. Lorrys are of the second categories slightly less prone to accidents and so on.

	2005	2006	2007	2008	2009
Two wheeler	154	174	159	191	191
Lorry	108	120	113	104	103
MTC Bus	71	74	66	79	74
Car	60	72	75	75	77
Van	63	57	59	52	38
Auto	48	48	40	41	36
UKV	35	37	36	46	40
Others	20	12	5	4	12
Private bus	14	22	15	22	19
Govt. bus	10	7	5	15	9

3. **Similarly vehicle types are classified based on non fatal injuries on accidents**

	2005	2006	2007	2008	2009
Two wheeler	1501	1370	1373	1517	1438
Car	903	966	1065	1694	1229
Auto	720	695	669	664	493
Lorry	473	425	375	654	456
Van	433	384	359	590	442
MTC Bus	184	202	274	334	285
Private Bus	63	59	61	129	68
Others	59	55	21	59	50
UKV	44	47	40	46	62
Jeep	29	23	22	26	29
Govt. bus	18	16	18	38	23

4.Number of deaths as per the victim and as per the death:

	2005	2006	2007	2008	2009
PEDESTRAIN	220	247	222	231	242
MCRIDER	148	200	211	236	202
CYCLIST	74	65	51	48	50
MCPRIDER	45	35	41	36	43
OTHERS	16	11	12	7	14
AUTODRIVER	11	6	9	7	4
AUTOOCCUPANT	9	17	13	7	7

5.Number of injuries as per the victim

	2005	2006	2007	2008	2009
MCRIDER	1741	1785	1907	1927	1743
PEDESTRAIN	1418	1335	1297	1373	1270
MCRIDER	1741	1785	1907	1927	1743
CYCLIST	493	407	347	347	241
AUTOOCCUPANT	314	251	301	275	208
AUTODRIVER	204	178	180	183	150
STANDPER	172	147	185	170	129
CAROCCUPANT	91	83	104	146	107
CARDRIVER	102	83	97	121	120
LORRYDRIVER	24	15	19	16	15
VANOCCUPANT	51	23	25	31	29
VANDRIVER	33	28	19	23	19

6. Number of fatal injuries as per the age for fatal male:

AGE	2005	2006	2007	2008	2009
15 to 29	130	149	182	169	11
30 to 44	131	139	117	141	116
45 to 59	127	147	107	132	127
ABOVE60	88	89	83	81	99
BELOW14	16	15	12	7	11

10. Number of fatal injuries based on the road:

ROAD	2005	2006	2007	2008	2009
100FeetRoad	55	38	34	43	43
OMR Road	38	35	33	52	44
ECR Road	47	51	30	28	40
AnnaSalai	35	42	32	38	33
ArcotRoad	23	17	16	26	16
200FeetRoad	14	24	15	16	15
ThiruvotriyurHighRoad	18	0	0	0	103
SPRoad	11	6	9	11	8
Tharamani Road	9	7	13	8	6
VelacherryMainRoad	8	14	14	18	11
SNChettyStreet	10	12	4	11	8
EnnoreExpressRoad	5	9	8	12	12
DurgabaiDeshmulkRoad	5	6	6	4	4
NewAvadiRoad	4	6	11	6	4
PoonamalleeHighRoad	0	9	3	2	4

7. Number of non-fatal injuries as per the age of male:

AGE	2005	2006	2007	2008	2009
15 to 29	1387	1364	1440	1398	1217
30 to 44	1254	1180	1170	1241	1044
45 to 59	793	695	765	783	713
ABOVE60	258	255	282	287	272
BELOW14	219	196	213	193	184

8. Number of fatal injuries as per the age of female:

AGE	2005	2006	2007	2008	2009
45to59	32	33	18	25	127
ABOVE60	28	21	28	41	29
30to44	20	18	23	13	17
15to29	17	18	17	12	16
BELOW14	7	10	10	6	6

9. Number of non-fatal injuries as per the age of female:

AGE	2005	2006	2007	2008	2009
30to44	306	283	289	291	286
15to29	287	253	242	278	210
45to59	229	200	255	240	215
ABOVE60	128	129	130	165	137
BELOW14	101	101	86	96	102

A Study of Voice over Internet Protocol

Mohsen Gerami

The Faculty of Applied Science of Post and Communications
Danesh Blv, Jenah Ave, Azadi Sqr, Tehran, Iran.
Postal code: 1391637111
e-mail: artimes0@hotmail.com

Abstract—**Voice over Internet Protocol, is an application that enables data packet networks to transport real time voice traffic. VOIP uses the Internet as the transmission network. This paper describes VoIP and its requirements. The paper further discusses various VoIP protocol, security and its market.**

Keywords: VOIP; H.323; SIP; Security; Market;

I. INTRODUCTION

VoIP, or Voice over IP, is an application that enables data packet networks to transport real time voice traffic. It consists of hardware and software that allows companies and persons to engage in telephone conversations over data networks. As a result, more and more companies have become interested in implementing VoIP [1].

All VOIP services are not built alike. Some allow you to call anyone with a phone, while others restrict your calls to only other clients using the same VOIP service. You can choose between three different ways to set up a VOIP system on your computer. You can use an ATA (analog voice adaptor) which performs the analog-to-digital conversion, and is plugged in to your computer at one end and your telephone at the other. You can use an IP phone, a phone specifically made for use with VOIP. While the IP phone looks exactly like a normal phone, it's got special Ethernet connectors that allow it to be plugged into your router. They're even working on WIFI phones for VOIP that you can take with you to the various internet hotspots popping up all over the world. Finally, you can make VOIP contact with your computer alone. Simply install the VOIP software, make sure you've got a microphone, speakers, an internet connection (high-speed is best, of course), and a sound card, and chat away. One thing many VOIP-users love about it is the cost, or more accurately, the savings. By using VOIP you save yourself one unnecessary bill per month - your phone bill. VOIP charges, much cheaper usually than most people's phone bills, appear on your regular broadband bill [2].

The technology underpinning VoIP was initially developed in the late 1970s, but it took almost 20 years to evolve from a computer novelty into a household service. It's now used by hundreds of thousands of people every day.

VoIP works in a relatively simple way. Each time you make a phone call your voice is converted into a stream of data. Then, rather than being sent over the phone network, this data stream travels over your broadband internet connection.

Each data packet is labelled with its destination address (the person you're calling) and moves through the internet in the same way as web pages and file downloads. When they get to their destination, the packets are reassembled and converted back into sound waves. When you have this process happening simultaneously in two directions, you've got a phone call.

Most VoIP services also come with an allocated landline phone number which allows other people to call you. In these cases the call will be routed to the nearest handover point (called a POP or point of presence) and then travel over the internet to your VoIP phone or computer [3].

II. REQUIREMENTS FOR VOIP

Obviously, the most important requirement is a broadband internet connection. Broadband connections are provided by cable companies (digital cable service), telephone companies (DSL, T1, etc.), and radio/microwave broadband internet connections. Currently, satellite (ie., satellite uplink dish) internet connections are not compatible with VOIP equipment because of the proprietary data compression algorithms used in satellite uplink and downlink. Further, the speed of light delay to and from a geosynchronous orbiting satellite would prove to be very annoying people trying to talk.

Broadband connection data uplink and downlink speeds of greater than 80 kilobits per second per telephone circuit (while a call is in progress) are generally considered to be the minimum requirement for "decent" voice transmission quality.

A "Telephone Adapter" (or "TA," and also known as an "Analog Telephone Adapter" or "ATA") is a piece of hardware that is used to digitize the voice and establish the IP session to the internet phone company's network switch. While it is possible to use a computer's microphone and speakers and special software for telephony over the Internet, the obvious limitation that the computer has to be turned on to make or receive phone calls makes this unwieldy.

A TA eliminates the need for a computer to be up and running and accepts a standard 4-wire RJ11 telephone cable to

support either premise telephone wiring or a direct connection of a standard analog telephone.

In addition, the TA usually includes a built in "router" that provides firewall isolation for the computers connecting to the Internet, as well as a Local Area Network (LAN) switch or hub. This allows efficient Internet connection sharing. The internet phone company usually provides or rent the TA, or they can be purchased at retail for a reasonable price [4].

The third requirement is a VoIP Service Provider (VSP) also known as an Internet Telephony Service Provider (ITSP). The Provider will supply you with an account and some form of "Telephone Number" [5].

The final requirement is one or two common variety analog telephone handset. Almost all commonly available wireless telephones and most two line phone sets will work with VOIP [4].

A. All together

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file for "MSW A4 format". The TA is usually a cable or DSL router that connects to the cable company's or DSL provider's supplied terminal (or "modem"). The customer's computer is then connected to the TA, as are the one or two standard (RJ11) telephone cables, which connect to either a wall outlet or a standard analog telephone, depending on whether telephone extensions are present or not. More than one computer usually can be connected to the TA to create a Local Area Network (LAN).
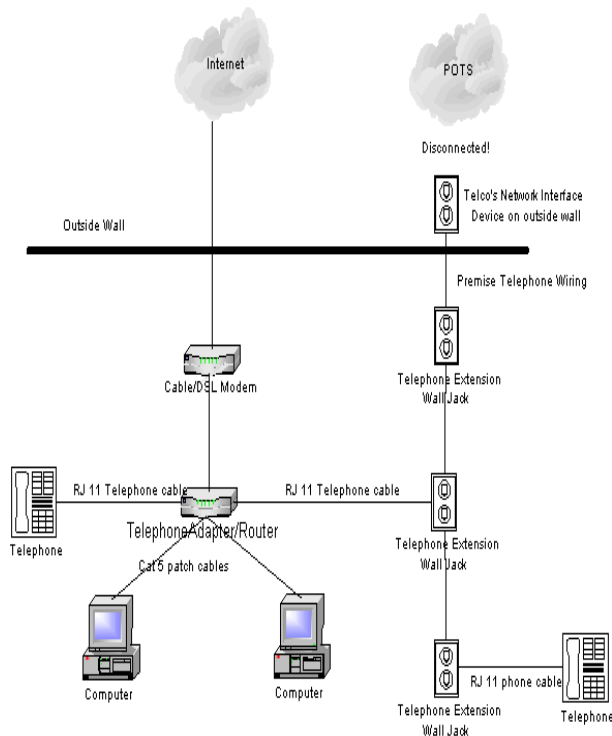

Figure 1.   A sample installation of VOIP.

The above diagram illustrates a sample installation. The telephone on the left is directly connected to the TA, while the one on the right is connected to the premise distribution wiring which is connected to the TA via RJ 11 cable to a wall jack.

The key to making VOIP work is the correct initialization of the TA to work with the internet phone company and configuring IP addresses on the router to use for the computer or premise LAN connected computers sharing the broadband connection [6].

III.   VOIP PROTOCOLS

To deliver voice, two types of VOIP protocol used: H.323 and SIP. H.323 and SIP both support VoIP and multimedia communications. H.323 is an older standard developed by the ITU. A good chunk of it is based on ISDN which comes from the traditional telephony world. H.323 is a binary protocol and is fairly complex in nature. SIP was developed by the Internet Engineering Task Force (IETF) and is text based (similar to HTTP). Much of the infrastructure already in place to support HTTP has been adapted to support SIP. IT managers within businesses are generally more comfortable with SIP because they are used to handling HTTP traffic. SIP is an open standard and solutions based on SIP are highly interoperable. A lot of effort has gone into ensuring interoperability and many manufacturers work together to regularly test to ensure this. Very few manufacturers are working on new H.323 implementations. SIP has become the standard of choice and is being worked on by large companies such as Microsoft and Cisco [7].

A.   H.323 Protocol Overview

H.323 is a ITU recommendation based on the H.320 family of standards. The current version of the recommendation is version 4 [8]. Initially, the protocol (version 1) was designed to provide signalling for a multimedia conferencing system for LAN environments with no quality of service provisions. However, in is current state, it has evolved into an umbrella of specifications that define the complete architecture and operation of a multimedia conferencing system over a wide area packet network. In contrast to its original scope, it has become a scalable solution that can be interworked with managed large scale networks.

A H.323 system provides the necessary signalling and control operations for performing multimedia communications over an underlying packet based network which may not provide a guaranteed quality of service. The actual network interface, the physical network and the transport protocols used on the network are not included in the scope of H.323. A H.323 system comprises of the following entities: Terminals, Gatekeepers, Gateways, Multipoint Controllers, Multipoint Processors and Multipoint Control Units.

• *Terminals* provide the audio/video/data communications capability in point-to-point or multipoint conferences, as well as handling the H.323 signalling issues on behalf of the user.

• *Gatekeepers* provide admission control and address translation services

• *Gateways* are needed to provide interworking with terminals using other signalling protocols, such as PSTN terminals, ISDN terminals, SIP terminals, etc.

• *Multipoint Controllers, Multipoint Processors and Multipoint Control Units* provide support for multipoint conferences.

A central aspect of H.323 is the *H.323 call.* It is defined as the point-to-point multimedia communication between two H.323 endpoints. If the H.323 endpoint communicates with an endpoint which uses a different signalling protocol, then the H.323 call is defined as the call segment between the H.323 entity and the gateway that provides interworking with the foreign network.

The H.323 protocol is a tightly coupled family of sub protocols which must all interoperate in order to complete successfully a multimedia call session. The sub protocols are described in ITU recommendations. The main ones are:

• *H.225:* Sub protocol for messages exchanged between H.323 endpoints for setting up and tearing down a call as well as for messages between an H.323 endpoint and its controlling H.323 entity, such as a gatekeeper.

• *H.245:* Sub protocol for messages exchanged between endpoints in order to control the call session, exchange resource capabilities and establish media channels.

• *H.235:* Sub protocol for security and encryption for H.323 terminals.

• *H.450:* Sub protocols for supplementary services, such as Call Transfer, Call Park, Call Waiting etc [9].

B. SIP Protocol Overview

SIP, which stands for Session Initiation Protocol, is an IETF application layer control protocol, defined in RFC 2543 [10], for the establishment, modification and termination of multimedia sessions with one or more participants. SIP makes minimal assumptions about the underlying transport and network layer protocol, which can provide either a packet or byte stream service with either reliable or unreliable service.

A SIP system is based on a client/server model and is comprised of the following logical entities:

• A *User Agent (UA)* is an application that acts on behalf of the user, both as a client (User Agent Client) and as a server (User Agent Server). As a client it initiates SIP requests and as a server it accepts calls and responds to SIP requests made by other entities. The user agent is usually part of a multimedia terminal whose media capabilities it controls without having any media capabilities of its own.

• A *Registrar Server* is a SIP server that accepts only registration requests issued by user agents. A registrar server never forwards requests.

• A *Location Server* is a server which provides information to a proxy/redirect server about the possible current locations of a user. Usually, this entity is part of the proxy/redirect servers.

• A *Redirect Server* is a SIP server that provides address mapping services. It responds to a SIP request destined to an address with a list of new addresses. A redirect server doesn't accept calls, doesn't forward requests nor does it initiate any of its own.

• A *Proxy Server* is a SIP server that acts both as a server to user agents by forwarding SIP requests and as a client to other SIP servers by submitting the forwarded requests to them on behalf of user agents or proxy servers.

With the exception of the user agent, which is usually part of a multimedia terminal, the rest of the logical entities (registrar, redirect and proxy servers)a may be combined in a single application. Therefore, a single entity can act either as a proxy or as a redirect server, according to the SIP request, and at the same time accept registration requests. A SIP call is defined as the multimedia conference consisting of all participants invited by a common source.

Although not partitioned formally, the SIP system can be viewed as divided into domains each serviced by one redirect/proxy server and one registrar. A user agent has usually a home domain, which is specified by its address, but it can roam and use services in other domains as well, in which case it is considered to be 'visiting'. Otherwise it is considered to be "at home" [9]

C. Related Work- Comparison of two Protocols

The authors of Nortel Networks [11] conclude by recommending SIP as their preference for a control protocol. They point out that even though H.323, unlike SIP, has currently more enterprise oriented and campus scale products deployed, SIP provides long term benefits which are related to and affect time to market, extensibility, multi-party service flexibility, ease of interoperability and complexity of development.

The Dalgic and Fang [12] concluded that In terms of functionality and services that can be supported, H.323v3 and SIP are very similar. However, supplementary services in H.323 are more rigorously defined and therefore fewer interoperability issues are expected to arise. Furthermore, H.323 has better compatibility among its different versions and better interoperability with the PSTN. The two protocls are comparable in their QoS support (similar call setup delays, no support for resource reservation or class of service (QoS) setting), but H.323v3 will allow signaling of the requested QoS. On the other hand, according to the paper, SIP's primary advantages are its flexibility to add new features and its relative ease of implementation and debugging. Finally, the authors note that H.323 and SIP are improving themselves by learning from each other, and the differences between them are diminishing with each new version.

The Schulzrinne and Rosenberg [13] wrote that SIP provides a similar set of services to H.323, but provides far lower complexity, rich extensibility, and better scalability. They point out that future work is due to more fully evaluate the protocols, and examine quantitative performance metrics to characterize these differences. They also imply that a study

measuring the processing overhead of SIP and H.323, would be quite useful [14].

IV. SECURITY

A future expectation is that long-established security features (i.e., authentication and encryption) will be integrated into VoIP standards. However, today many existing data-centric security technologies can be utilized to enhance security in the VoIP environment. VoIP network security includes voice-packet security, which focuses on application concerns, while IP security focuses on transport or network security. Controlling security at these levels of the VoIP environment may require network re-design and/or re-engineering which will affect the architecture of the network supporting the VoIP environment. Some specific issues need further attention when a VoIP system is deployed. It is important to remember that securing any network is a continual process that requires staying abreast of the latest vulnerabilities that may exist in network infrastructure components, server operating systems, and applications deployed throughout the enterprise [15].

In the early days of VoIP, there was no big concern about security issues related to its use. People were mostly concerned with its cost, functionality and reliability. Now that VoIP is gaining wide acceptance and becoming one of the mainstream communication technologies, security has become a major issue.

The security threats cause even more concern when we think that VoIP is in fact replacing the oldest and most secure communication system the world ever known – POTS (Plain Old Telephone System).

A. Security Threats in VoIP

Service theft can be exemplified by phreaking, which is a type of hacking that steals service from a service provider, or use service while passing the cost to another person. Encryption is not very common in SIP, which controls authentication over VoIP calls, so user credentials are vulnerable to theft.

Eavesdropping is how most hackers steal credentials and other information. Through eavesdropping, a third party can obtain names, password and phone numbers, allowing them to gain control over voicemail, calling plan, call forwarding and billing information. This subsequently leads to service theft.

Stealing credentials to make calls without paying is not the only reason behind identity theft. Many people do it to get important information like business data.

A phreaker can change calling plans and packages and add more credit or make calls using the victim's account. He can of course as well access confidential elements like voice mail, do personal things like change a call forwarding number.

Vishing

Vishing is another word for VoIP Phishing, which involves a party calling you faking a trustworthy organization (e.g. your bank) and requesting confidential and often critical information. Here is how you can avoid being a vishing victim.

Viruses and malware

VoIP utilization involving softphones and software are vulnerable to worms, viruses and malware, just like any Internet application. Since these softphone applications run on user systems like PCs and PDAs, they are exposed and vulnerable to malicious code attacks in voice applications.

DoS (Denial of Service)

A DoS attack is an attack on a network or device denying it of a service or connectivity. It can be done by consuming its bandwidth or overloading the network or the device's internal resources.

In VoIP, DoS attacks can be carried out by flooding a target with unnecessary SIP call-signaling messages, thereby degrading the service. This causes calls to drop prematurely and halts call processing.

Why would someone launch a DoS attack? Once the target is denied of the service and ceases operating, the attacker can get remote control of the administrative facilities of the system.

SPIT (Spamming over Internet Telephony)

If you use email regularly, then you must know what spamming is. Put simply, spamming is actually sending emails to people against their will. These emails consist mainly of online sales calls. Spamming in VoIP is not very common yet, but is starting to be, especially with the emergence of VoIP as an industrial tool.

Every VoIP account has an associated IP address. It is easy for spammers to send their messages (voicemails) to thousands of IP addresses. Voicemailing as a result will suffer. With spamming, voicemails will be clogged and more space as well as better voicemail management tools will be required. Moreover, spam messages can carry viruses and spyware along with them.

This brings us to another flavor of SPIT, which is phishing over VoIP. Phishing attacks consist of sending a voicemail to a person, masquerading it with information from a party trustworthy to the receiver, like a bank or online paying service, making him think he is safe. The voicemail usually asks for confidential data like passwords or credit card numbers. You can imagine the rest!

Call tampering

Call tampering is an attack which involves tampering a phone call in progress. For example, the attacker can simply spoil the quality of the call by injecting noise packets in the communication stream. He can also withhold the delivery of packets so that the communication becomes spotty and the participants encounter long periods of silence during the call.

Man-in-the-middle attacks

VoIP is particularly vulnerable to man-in-the-middle attacks, in which the attacker intercepts call-signaling SIP message traffic and masquerades as the calling party to the called party, or vice versa. Once the attacker has gained this position, he can hijack calls via a redirection server [16].

What Affects Voice Quality in VoIP Calls

Here are the main things that affect voice quality in VoIP and what can be done to maximize quality.

Bandwidth

Your Internet connection always tops the list of factors affecting voice quality in VoIP conversations. The bandwidth you have for VoIP is the key for voice quality. For instance, if you have dial-up connection, don't expect great quality. A broadband connection will work right, as long as it is not spotty, and not shared with too many other communication applications. Bandwidth dependency is besides one of the main drawbacks of VoIP.

Equipment

The VoIP hardware equipment you use can greatly impact on your quality. Poor quality equipment are normally the cheapest ones (but not always!). It is therefore always good to have as much information as possible on an ATA, router or IP phone before investing on it and starting to use it. Read reviews and discuss about it in forums. It might also be that the hardware you choose is the best in the world, but still you get problems - because you are not using hardware that suits your needs.

ATA/Router

For an ATA/Router, you need to think of the following:

• Compression technologies (codecs) supported

• Echo cancellation, which is a mechanism for decreasing echo

• Firewall and security support

Phone frequencies

The frequency of your IP phone may cause interference with other VoIP equipment. There are many cases where people using 5.8 GHz phones have been getting voice quality problems. When all troubleshooting tricks failed, changing the phone to one with a lower frequency (e.g. 2.4 GHz) solved the problem.

Weather Conditions

At times, the voice is terribly distorted by something called static, which is a small 'dirty-weed' static electricity generated on broadband lines due to thunderstorms, heavy rain, strong gusts, electrical impulses etc. This static is not very much noticeable when you surf the net or download files, which is why we don't complain about it when we use the Internet for data despite it be here; but when you are listening to voice, it becomes disturbing. It is easy to get rid of static: unplug your hardware (ATA, router or phone) and plug it back again. The static will be brought to naught.

The effect of weather conditions on your connection is not something you can change. You can have some short-term relief in some cases, but most of the time, it is up to your service provider to do something. At times, changing the cables solves the problem completely, but this can be costly.

Location of your hardware

Interference is a poison for voice quality during voice communication. Often, VoIP equipment interfere with each other thus producing noise and other problems. For example, if your ATA is too close to your broadband router, you might experience voice quality problems. This is caused by electrical feedback. Try moving them away from each other to get rid of the garbled calls, echoes, dropped calls etc.

Compression: the codec used

VoIP transmits voice data packets in a compressed form, so that the load to be transmitted is lighter. The compression software used for this are called codec's. Some codecs are good while others are less good. Put simply, each codec is designed for a specific use. If a codec is used for a communication need other than that for which it is meant, quality will suffer. [17]

V. VoIP MARKET

Telefonica has acquired VoIP provider Jajah, continuing the consolidation trend in the sector. Microsoft already bought Tellme, BT acquired Ribbit, Google took over GrandCentral and KPN is buying out the minority shareholders in iBasis. In addition, eBay sold a majority stake in Skype to Silver Lake Partners, which also has a stake in Avaya, while Skype and Avaya have started talks on working together. Another hardware manufacturer, Nortel, has received an early bid for its VoIP assets, from Genband.

The trend shows newcomers slowly but surly losing their independence by joining larger groups. The latest takeover, Telefonica's acquisition of Jajah, is perhaps the most remarkable in that sense, as Jajah was originally set up to avoid the high international tariffs charged by the incumbents. However, Google and Skype are still maintaining their independence.

There are still plenty of potential acquisition targets, at various stages of the VoIP value chain: iSkoot, Truphone, Jaxtr, Fring, Nimbuzz, Ooma, Vonage, 8x8 (Packet8), Rebtel, Freshtel, Mobivox, Sipgate, Vyke, Telio, Snapvine and many others.

All these VoIP services providers combine infrastructure with services provision for end-users, and the question is what the new owners will do with acquired assets. The choice comes roughly down to wholesale (capacity, platform services, software) and retail (VoIP services for end-users). BT (Ribbit) and KPN (iBasis) are choosing clearly for the wholesale side, and that seems to also be the case with Telefonica (Jajah). Operators as well as large corporations can be offered VoIP services based on the acquired infrastructure, platforms and software.

Google and Skype represent, as 'newcomers' on the telephony market, the retail side. As for Jajah, the question is whether Telefonica will maintain the end-user services, or slowly dissolve these in order to protect its own international business. There is a clear reason to keep this side of the Jajah business though: whatever Telefonica is losing on its home market, it can win back abroad by competing with incumbents elsewhere for international business.

That still leaves the question of what's the future for newcomers such as Google and Skype. Pressure from VoIP will eventually drive all call costs to the price of local calls. After deducting termination fees, there is little left over for the provider. Call services will then become a true commodity business. That's the doomsday scenario for Skype, as it mainly makes money from avoiding high international call tariffs. Skype will need to build up quickly a large customer base, in order to offset the margin erosion. For Google, this is less of a problem, as it already has various services that do not directly earn money. For the incumbents, which are already seeing a sharp decline in their international business, there is a need to move further up the value chain. Companies such as BT and Telefonica can be expected to drive innovation on the telephony market going forward, with especially business customers profiting. Skype may also, in cooperation with Avaya target the business market, in order to exploit new income sources [18].

A. France led VoIP market in Europe in Q309

With more than 15 million subscribers, France is leading the Voice over IP (VoIP) market in Europe. Orange France is the telecom operator in Europe with the highest VoIP subscriber base with 6.580 million subscribers, it represents 17% of the European VoIP market.

According to Dataxis Intelligence, in Q309 there were 39.7 million VoIP subscribers. This figure includes Voice over DSL, Cable and Fiber. The split is 76% over DSL, 21.8% over Cable and 2.2% over Fiber [19].
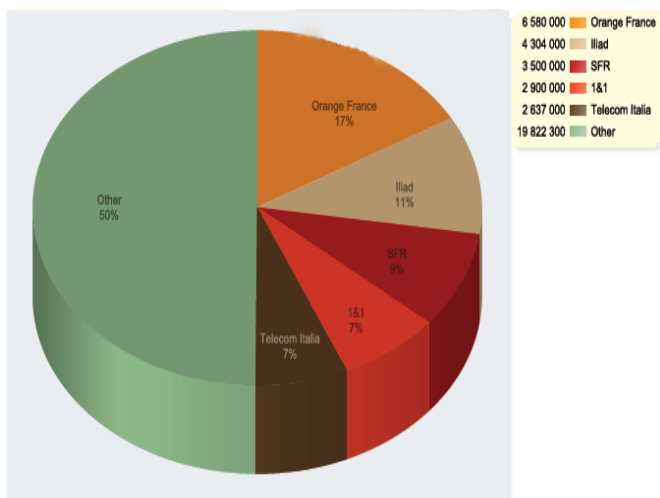
Figure 2.   VoIP market in Europe in Q309- Source : *Dataxis Intelligence*

B. SERVICE PROVIDER VOIP AND IMS

"The worldwide carrier voice over IP equipment market closed the 2009 calendar year down 28%, as expected, ending with its third consecutive quarter of stable revenue in the fourth quarter, led by very strong session border controller sales. Meanwhile, the IMS equipment market ended on a high note with 2009 worldwide revenue up 142% over 2008. The shining star in 4Q09 for IMS sales were deployments for mobile networks, particularly purchases for Rich Communication

Suite, video, LTE trials and enhanced mobile IM and presence services. The IMS market will continue to be lumpy on a quarterly basis, but we expect continued positive momentum from new deployments from North American cable operators, Class 5 replacement projects in EMEA and VoLTE to contribute to strong annual growth for at least the next five years," forecasts Diane Myers, directing analyst for service provider VoIP and IMS at Infonetics Research.

C. IMS MARKET HIGHLIGHTS

- First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file for "MSW A4 format". Worldwide IP Multimedia Subsystem (IMS) equipment vendor revenue totaled $426 million in 2009 and is forecast to grow to $1.44 billion in 2014

- Following a down 3Q09, the worldwide IMS equipment market posted its strongest quarter to date, jumping 92% sequentially in 4Q09

- 4Q09 marked the first quarter in which revenue from IMS equipment for mobile networks surpassed that of IMS equipment for fixed-line networks

- Alcatel-Lucent and Nokia Siemens Networks each posted very strong IMS equipment results in 4Q09

- With key operators and vendors forming the OneVoice initiative and transferring the initiative to the GSMA in February 2010, IMS is guaranteed to get its biggest driver from LTE deployments starting in 2012

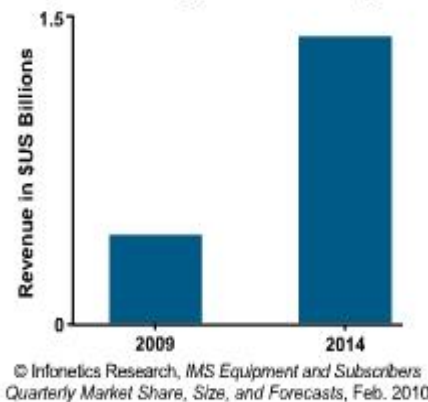Figure 3.   Worldwide IMS growth.

D. SERVICE PROVIDER VOIP MARKET HIGHLIGHTS

- The worldwide service provider VoIP equipment market dropped 28.2% from 2008 to 2009, to $2.48 billion worldwide

- The migration to native IP sped up during the past 12 months due to the acceleration of TDM access line loss, resulting in significant declines in TDM-related equipment, particularly traditional trunk media gateways and softswitches

- In 4Q09, service provider VoIP equipment revenue was up slightly, at 2.7% over 3Q09, continuing the period of relative stability that started in 2Q09

- Four vendors stood out for growing revenue in 2009:
    o Metaswitch in trunk media gateways and softswitches
    o Acme Packet in session border controllers
    o Radisys in media servers
    o BroadSoft in voice application servers

- GENBAND's pending acquisition of Nortel's CVAS unit will cause some significant shifts in the vendor landscape, making GENBAND the largest carrier VoIP vendor in terms of overall revenue [20].
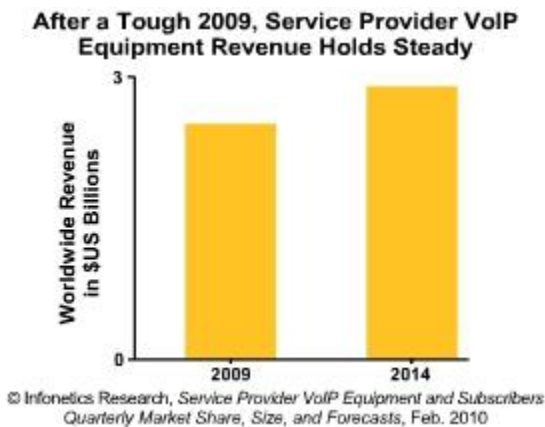
Figure 4. Worldwide Service Provider IMS growth.

VI. CONCLUSION

VoIP is the voice of the future however many problems still remain.. VoIP phones are significantly less expensive than traditional phone lines. VoIP companies need some methods to make revenue, and they need a unique business plan; therefore demands have to rise. A VOIP user can call any other user, located anywhere in the world, with better voice quality. That is ideal for VOIP service. For next step the equipment will be more acceptable and technology has to present more power and security enhances user's trust.

REFERENCES

[1] http://www.comtest.com/tutorials/VoIP.html

[2] http://articles.directorym.com/Importance_Of_VOIP-a971521.html

[3] Ian Grayson, 2009, www.cnet.com.au/voip-guide-voice-over-ip-in-australia-240056481.htm

[4] http://www.zoesnet.net/VOIP.htm

[5] http://www.thevoipstore.net/VoIP-Requirements.php

[6] http://www.zoesnet.net/VOIP.htm

[7] http://searchunifiedcommunications.techtarget.com/expert/Knowledgebas seAnswer/0,289625,sid186_gci1069115,00.html

[8] I. T. Union, "Packet-based multimedia communication systems," Telecommunication Standardization Sector of ITU, Geneva, Switzerland, Recommendation H.323, Nov. 17, 2000.

[9] Papageorgiou Pavlos ,2001, A Comparison of H.323 vs SIP

[10] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, "SIP: Session Initiation Protocol," Internet Engineering Task Force, Request For Comments 2543, Mar. 1999, proposed Standard.

[11] N. Networks, "A Comparison of H.323v4 and SIP," 3GPP S2, Tokyo, Japan, Technical Report S2-000505, Jan. 5 2000.

[12] I. Dalgic and H. Fang, "Comparison of H.323 and SIP for IP Telephony Signaling," in *Proceedings of SPIE. Multimedia Systems and Applications II*, ser. Proceedings of Photonics East, Tescher, Vasudev, Bove, and Derryberry, Eds., vol. 3845. Boston, Massachusetts. USA: The Internationl Society for Optical Engineering (SPIE), Sept. 20-22 1999.

[13] H. Schulzrinne and J. Rosenberg, "A Comparison of SIP and H.323 for Internet Telephony," in *Proceedings ofThe 8th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 98)*, Cambridge, UK, July 8-10 1998, pp. 83–86.

[14] Papageorgiou Pavlos, 2001, A Comparison of H.323 vs SIP, University of Maryland at College Park

[15] DISA for the DOD, 2006, Internet Protocol Telephony & Voice Over Internet Protocol, Version 2, Release 2

[16] Nadeem Unuth, http://voip.about.com/od/security/a/SecuThreats.htm

[17] Nadeem Unuth, http://voip.about.com/od/voipbasics/a/factorsquality.htm

[18] telecompaper, 2009/12/30, http://en.c114.net/583/a472219.html

[19] Dataxis Intelligence, 2010, http://dataxisnews.com/?p=11055

[20] Infonetics Research Report Highlights, 2010, http://www2.marketwire.com/mw/mmframe?prid=593499&attachid=11 91085

Performance Issues of Health Care System using SQL server

Narendra Kohli

Electrical Engineering Department

Indian Institute of Technology

Kanpur, India

nkohli@iitk.ac.in

Nishchal K. Verma

Electrical Engineering Department

Indian Institute of Technology

Kanpur, India

nishchal@iitk.ac.in

Abstract—: **In this paper, a smart card based on line health care system and its performance issues using SQL server are proposed. To provide a good quality of treatment in the hospital, it is required to integrate all the hospitals of country via internet. A Smart Card with 10 digits unique registration no. with his some personal information is issued to patient. After getting registration in any hospital of the hospital network, patient has to go for checkup with smart card only. All the patient information i.e. personal, doctor prescriptions, test reports etc. will be stored in the database of the local server of the hospital and time to time uploaded to the centralized server. On the basis of unique registration no., all the patient information can be retrieved from the database of the centralized server. Smart card based online health care system application has been designed as front end .Net and back end in SQL server. The block size or page size being used during the database creation is playing very important role in performance tuning. It is very important to decide the proper block size before database design. You cannot change the block size once you have created the database. Re-creating the database again is a very costly affair.**

Keywords- hospital, patient, smart card, SQL server 2005

I. INTRODUCTION

Automation & networking of the hospitals are the necessity of the society. The purpose of the same is to provide the better services for patients and it will increase the working efficiency of the hospital system. Main idea for health care system is to obtain, store, analysis or process and uses of patient information (patient, doctor, hospitals, laboratory tests etc.). At registration counter in a hospital, administrator will generate a 10 digit unique patient-Id. Basic information's i.e. name, address, phone no. etc. with unique patient id will be stored on the smart card and issued to the patient. As per the patient id, patient related information i.e. doctor diagnosis, test reports, MRI, CT-scan images etc will be stored in the databases of the hospital server and time to time uploaded to the centralized server. In future this information will be useful for doctors to diagnose the illness and give the important suggestions to the patients for their health. While visiting to the hospital for treatment, patient has to carry only smart card [1] [2]. Administrator of the hospital or doctor will use the smart card

through card reader. Designing of the proper databases and uses of different indexing techniques in SQL server 2005 will help for fast retrieval of patient data.

II. Literature review

The need for automation systems in hospitals gains more importance [4]. A patient may be registered in any hospital of the networking of the hospitals. As per the requirement it is important to share the patient information between different hospitals without ignoring privacy constraints & other related information. [1] Proposed the use of smart card to store the basic information of the patient. Nowadays different types of smart cards are exiting in market [2] [3]. These cards have proven to be convenient tokens for identification and authentication in day to day activities [5]. As per the requirement, different types of smart cards can be used to store the patient information. [6] [7] explains about telemedicine and PACS. E-health is producing a great impact in the field of information distribution of the health services to the hospitals and public [8] [9]. Query optimization of SQL server has been discussed in [10] [11].

III. PROBLEM FORMULATION

3.1   If a patient is registered in one hospital and integration and retrieval of patient information is not possible in the hospital system then while taking the consultation with doctor, patient can easily forget to explain his previous treatment. In case of which incorrect prescriptions may be applied. Keeping in mind an intelligent system i.e. smart card based online health care system has been proposed. The proposed system is given in Fig. 1. The servers for the hospitals with high technical configuration are required. All the servers of the hospitals are connected with one centralized server of the hospital through internet. A very high bandwidth dedicated internet lease line has been proposed to use for the system. A smart card reader / writer unit has been attached to each computer of this hospital system network. The proposed health care system has been loaded to all the servers of all the hospitals. The patient smart card stores some important

information like unique patient id, name, sex, date of birth, blood group etc. As per the patient-id, patient details like treatment prescriptions, test reports, images like MRI, CT-scan etc. have been stored in the database of the hospital server. On the basis of stored details of the patient, doctor can prescribe the proper medicine. SQL server tuning has been used while designing the application.
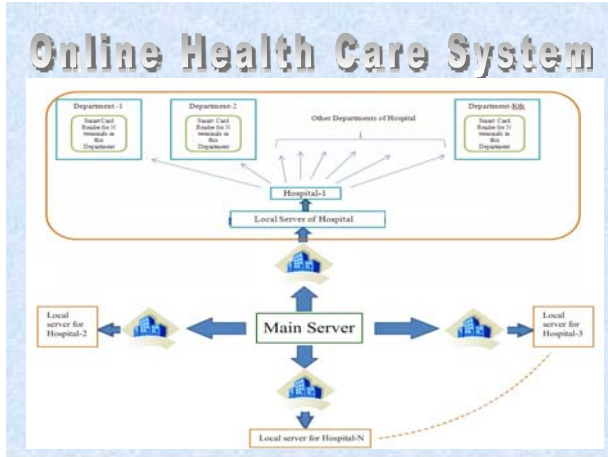

Fig. 1: Smart card based health care system

3.2 To increase the performance of the application of online health care system, the following issues & technical issues will be taken:

3.2.1 Issues:

- Design an application that can serve maximum number of hospital and provide a platform for their interaction.
- A system generated patient id that is unique for maximum possible time interval.
- Store patient information optimally.
- Speed up the DML (insert, update, delete) operations.
- Store patient information for maximum possible time.

3.2.2 Technical issues:

- Normalization of tables.
- Maximum utilization of memory.
- Imposing necessary constraints on tables.
- Selecting best possible structure for queries.
- Tuning and optimizing queries.
- Optimal selection of join order and join algorithm.

IV Methodology

Approach for above issues:

1. The patient will provide basic information through smart card (having 10 digit patient id and personal information) at registration counter where it will be uploaded to server.

2. In smart card we are storing patient id and personal information like name, address, DOB etc. of a person. As per the size of smart card we can store the information of a person.

3. Patient's files may be one of the two forms image or PDF. Images are stored separately in image tables and PDF in PDF table .These tables contain an Image ID or PDF ID and their description. Image ID and PDF ID are stored in their respective tables as foreign keys that may be in patient doctor or Patient Lab or in PatientRoom Tables.

4. Since we are storing all the patient related information's on a centralized database server. So movements of files is not required .The required files for a particular patient & their lab test report & images, prescriptions, scan document & diagnosis reports can be retrieved from centralized database server. Accessing the information's related to patient will be controlled by various level of access control and it will help us to prevent the unauthorized access from the data base.

5. As we have normalized our database we have removes all possible redundancies

6. We have covered optimization at application design level, database design level, memory utilization & optimization of queries for accessing the data base. Few points about the optimization are as follows:

- Application is studied well and requirements are identified. These requirements are categorized on object level and respective tables to store that information are created. After that database is normalized up to BCNF (Boyce Codd Normal Form) level and 14 tables are created.
- Size of attributes (their respective data types) is further minimized in order to compact record size, So that maximum number of records can be stored in a single data page. Select queries are designed to handle all possible search criteria.

- Indexes are created to speed up above read access. First data base ids well studied to find out what possible values different attributes will take and what will be the size of tables. Considering both the things Hash, B-tree (clustered and non clustered) and Bitmap indexes are proposed.

- Order of attributes in composite indexes are studied and justified.

- Every hospital will have day to day patient information locally to improve the performance

particularly in case of frequent insert, update operation.

- In a multi table join queries, we have studied the nature of the table & depending on the characteristic of the table their order will be pre estimated. SQL queries Hints can also be given to optimizer to follow that order to achieve the optimal performance.

- Proper join algorithms are justified in respective cases and hints are given to optimizer.

- Selecting the most optimal execution path by the optimizer is the time consuming process. By given the Hints we can reduce the time taken by the optimizer up to some level

- To improve the DML operation by the hospital we try to keep the data initially at the local server. This scheme reduces load from centralized server.

- Every day a particular time is selected to update central server by collecting information from local servers automatically. The operation to refresh the centralized server may be made more frequent if required.

- Transaction log is permanently off at centralized server as we don't need any log for recovery. Because the all information's are at the local server & can be reproduced easily.

- All primary key, foreign key constraints are removed from centralized server as those constraints are already checked at local servers. So there is not need to revalidate the information which is validated. It will improve the data loading performance at the central server.

- For updating centralized server we will recommend to use bulk copy and bulk insert schemes

V. Findings

5.1 Simulation of query optimization for performance tuning:

To identify the bottleneck and performance related issues we have done the following case study using different methods and analyze the query and their execution path. In our database out of 14 tables, 4 tables stores patient related information's and every table has approximately 14 lac records in SQL server to simulate.

- Patient Stores registration information of patient

- PatientDoctor Stores Doctor related information like disease and others

- PatientLab stores test report of patients.

- PatientRoom Stores Room id's, Daily charges And other attributes of room where Patient is admitted

The following analysis is based on query and stored procedure given below:

The query selects Patient details from four different tables by applying inner joins.

Query:
SELECT Patient1.PatName, Patient1.Address, Patient1.ContactNo, Patient1.Sex, Patient1.Occupation, Patient1.Guardian, Patient1.RegNo,Patient1.ReferredBy, Patient1.BloodGroup, Patient1.PatientHistory, Patient1.Disease, Patient1.DiseaseCatageory, Patient1.EntryDate,PatientRoomIPD.Hospital_Id, PatientRoomIPD.Room_Id, PatientRoomIPD.StartDate, PatientRoomIPD.EndDate,PatientRoomIPD.Status, PatientRoomIPD.TotalCharge, PatientDoctor.DoctorId, PatientDoctor.DoctorName, PatientDoctor.StartDate AS TreatmentStartDate,PatientDoctor.EndDate AS TreatmentEndDate, PatientDoctor.No_Of_Visits, PatientDoctor.Charges AS TreatmentCharge, PatientLab.LabNo,PatientLab.TestNo, PatientLab.TestDate,PatientLab.Unit
FROM Patient1
INNER JOIN
PatientRoomIPD ON Patient1.PatId = PatientRoomIPD.PatID
INNER JOIN
PatientDoctor ON Patient1.PatId = PatientDoctor.PatId
INNER JOIN
PatientLab ON Patient1.PatId = PatientLab.PatId

where Patient1.PatId=@Id;

On the hardware configuration: Intel® Duel Core 2.8GHz 512 RAM

Table 1: results on front end and query analyzer

| Iteration | On Asp.net Search Page (Front-End) | | On Query Analyzer | |
	Execution time in ms using Stored Procedure	Execution time in ms using Query	Execution time in ms using Stored Procedure	Execution time in ms using query
1	.719	.771	.827	.740
2	.750	.756	.827	.704
3	.750	.722	.720	.716
4	.813	.733	.766	.843
5	.719	.781	.810	.797
6	.844	.727	.784	.767
7	.672	.697	.750	.803
8	.765	.755	.780	.767
Mean	.754	.743	.783	.767

Since we are taking different ID's for different iterations that's why stored procedures are complied each time when we call the procedure. The optimizer and analyzer check the SQL statement line by line and character by character and it treats the query differently every time when we give different IDs. As per the table 1 stored procedures taking much time than a query because we are unable to take the benefits of precompiled stored procedure (they are compiled every time when they are called in our case because it has more intelligence to reuse the query)

5.2 Optimization based on memory management:

SQL Server 2005 stores data in a special structure called data pages that are 8Kb (8192 bytes) in size. Some space on the data pages is used to store system information, which leaves 8060 bytes to store user's data. So, if the table's row size is 4040 bytes, then only one row will be placed on each data page. If you can decrease the row size to 4030 bytes, you can store two rows within a single page. The less space used, the smaller the table and index, and the less the I/O SQL Server has to perform when reading data pages from disk. So, one should design the tables in such a way as to maximize the number of rows that can fit into one data page and you should specify the narrowest columns you can. The narrower the columns, the less data that is stored, and the faster SQL Server are able to read and write data. On the above discussion it is clear that row size of a table should be some devisor of 8060 to maximum utilizes memory and speed up read/write from database.

So table size should be one of the following magic numbers

1	2	4	5	10	13	20
	26	31	52	62	65	124
	130	155	260	310	403	620
	806	1612	2015	4030	8060 bytes	

So the modification of table row size are based on above magic number

New Databases:

We have executed query and stored procedure from front end and query analyzer by taking same Patient ID at random on new data base after modification of table row sizes. Results are as table 2

Table 2: results on modified table row sizes

Iteration	On Asp.net Search Page (Front end)		On Query Analyzer	
	Execution time in ms using Stored Procedure	Execution time in ms using Query	Execution time in ms using Stored Procedure	Execution time in ms using Query
1	.562	.625	.563	.610
2	.609	.625	.610	.580
3	.578	.593	.610	.593
4	.578	.625	.594	.610
5	.609	.609	.593	.597
6	.563	.625	.594	.577
7	.531	.547	.594	.593
8	.578	.610	.593	.577
Average	.576	.607	.593	.592

After optimization of page and block sizes in the databases average execution time is reduced as per table 2.

Time in stored procedure and query:

0.754 ms (Old value), 0.576 ms (new value)
0.743 ms (Old value), 0.607 ms (new value)

5.3 Effect of indexing:

An index is a database objects that, when a table is created, can provide faster access path to data and can facilitate faster query execution.

a) After indexing on PatID of PatientDoctor table, the results are as table 3.

Table 3: Indexing on PatId of PatientDoctor table

Iteration	On Asp.net Search Page(front end)		On Query Analyzer	
	Execution time in ms using Stored Procedure	Execution time in ms using query	Execution time in ms using Stored Procedure	Execution time in ms using Query
2	.297	.344	.326	.330
3	.344	.390	.390	.300
4	.297	.390	.356	.330
5	.313	.328	.300	.327
6	.266	.359	.374	.343
7	.344	.328	.330	.313
8	.359	.344	.373	.357
9	.328	.375	.326	.330
Average(ms)	.318	.357	.347	.329

b) After indexing on PatID of PatientDoctor and PatientLab table, results are as in table 4.

Table 4: Indexing on PatID of PatientDoctor and PatientLab table

Iteration	On Asp.net Search Page		On Query Analyzer	
	Execution time in ms using Stored Procedure	Execution time in ms using query	Execution time in ms using Stored Procedure	Execution time in ms using Query
1	.265	.312	.250	.250
2	.297	.297	.297	.297
3	.281	,281	.314	.283
4	.312	.250	.296	.310
5	.297	.297	.250	.326
6	.297	.250	.250	.280
7	.329	.297	.280	.250
8	.281	.297	.280	.250
Average	.295	.286	.277	.280

c) After indexing on PatID of PatientDoctor and PatientLab and Patient Room table with proper memory optimization, results are as in table 5.

Table 5: Indexing on PatID of PatientDoctor & PatientLab & PatientRoom table

Iteration	On Asp.net Search Page		On Query Analyzer	
	Execution time in ms using Stored Procedure	Execution time in ms using query	Execution time in ms using Stored Procedure	Execution time in ms using query
1	.000	.047	.077	.000
2	.000	.000	.080	.076
3	.031	.031	.047	.077
4	.031	.063	.064	.063
5	.031	.000	.077	.047
6	.016	.000	.034	.060
7	.000	.000	.140	.060
8	.063	.031	.047	.013
Average	.019	.017	.070	.049

d) After indexing on PatID of PatientDoctor and PatientLab and PatientRoom on old data base without memory optimization, we are getting poor results. Results are as in table 6

Table 6: Indexing on PatId of PatientDoctor & PatientLab & PatientRoom table without memory optimization.

Iteration	On Asp.net Search Page		On Query Analyzer	
	Execution time in ms using Stored Procedure	Execution time in ms using query	Execution time in ms using Stored Procedure	Execution time in ms using Query
1	.125	.110	.076	.110
2	.094	.094	.110	.140
3	.110	.078	.123	.093
4	.078	.157	.047	.144
5	.110	.063	.120	.047
6	.094	.094	.060	.153
7	.093	.094	.110	.080
8	.079	.031	.140	.126
Average	.098	.090	.098	.112

V. Conclusions:

- Stored procedure using front end has better understanding about the SQL statement. If we are changing only some value for a given ID then it understand that the SQL statement is exactly the same as the previously executed statement with different value only. So the optimizer by passes the parsing, query optimization and generation of new execution path, and used the already available path for the next statement with different ID. This conclude that instead of using anonymous SQL statement we should recommend to use the stored procedure which gets stored in the database in the compiles format to avoid re-compilation every time as in case of simple un-name SQL statement or anonymous SQL statement. As it's stored permanently in the database so it can be cached in the memory and can be shared by several applications and also improves the reusability.

- The block size or page size being used during the database creation is playing very important role in performance tuning. If we know the average row size for a table that before going to create the database we should calculate the block sizes in such a way that maximum number of records can be stored in one page. It will drastically reduce the number of I/Os means less number of disk activity will be performed to fetch the data. It is very important to decide the proper block size before database design. You cannot change the block size once you have created the database. Re-creating the database again is a very costly affair.

- After creating indexes on the four tables query execution and stored procedure execution is taking less time than before. This is a good result. But this conclusion is only based on the B-tree index which is

better for the column which has very high cardinality & uniform distribution of the data. We have further explored the possibilities of using different kind of indexing techniques in case if the cardinality is low, data is not uniformly distributed & indexing column is monotonically increasing number. In these cases B-tree indexes will reduce the performance instead of improvements .In my other work I have proposed to incorporate hash, bitmap & Btree indexing techniques in database engine so that the optimizer will follow the most optimal query execution path.

REFERENCES

[1]   Smart Card Handbook.
[2]   Advance Card systems: http://www.acs.com.hk/index.php.
[3]   http://www.parivahan.nic.in.
[4]   Narendra kohli, Nishchal K. Verma. Performance issues of smart card based online health care automation system. Proceedings of the 1st international conference on Signals, systems & automation. 28th -29th December. India. 2009.
[5]   Eugene Lockett, Sungkyu park,Gueng cheng Jiang,Mike riddle. Security aspects of smart cards-term project CS 574 Fall 2003 San Diego state university, Nov.3, 2003.
[6]   J.R.Campbell and R. Stoupa. The patient, the provider, the processor: information management in ambulatory care. proceeding SCAMC,ed. pp. 930-940 (IEEE Computer Society Press 1990)
[7]   J.R.Scherrer. the Hospital information system- integrated patient records. Vol 48. Elsevier Science Ireland Ltd. 1995.
[8]   Jinman Kim, David Dagan Feng, Tom Weidong Cai, Stefan Eberl. Integrated Multimedia Medical Data Agent in E-Health. Pan-Sydney Area Workshop on visual Information Processing .Sydney, Australia. 2001
[9]   L.L.weed. knowledge coupling: New premises and new tools for medical care and education .Springer Verlog. New York. 1991.
[10]   SQL server tutorials:sqlserverpedia.com/wiki/SQL-Server-Tutorial
[11]   SQL server tutorial: http:// www.quackit.com/sql_server/tutorial/.

AUTHORS PROFILE

Narendra Kohli is working as Assistant Professor in the computer science and engineering deptt. , HBTI Kanpur. He is doing research on Telemedicine and PACS at IIT Kanpur.

Nishchal K. Verma is working as Assistant professor in Electrical engineering deptt., IIT Kanpur. His research interests are Machine Learning, Biometrics, GMM, HMM, Fuzzy Systems, Clustering Algorithms, Coloe Segmentaion,Video Image sequence recognition.

Color Steel Plates Defect Detection Using Wavelet And Color Analysis

Ebrahim Abouei Mehrizi
Department of Electronic Engineering
Islamic Azad University, najafabad branch
Isfahan, 81746, Iran
E_Aboei@yahoo.com

Amirhassan Monadjemi
Department of Computer Engineering
University of Isfahan,
Isfahan, 81746, Iran
Monadjemi@eng.ui.ac.ir

Mohsen Ashorian
Department of Electronic Engineering
Islamic Azad University, shahremajlesi branch,
Isfahan, 81746, Iran
Mohsena@ieee.org

Abstract— **In this study, having reviewed the automatic surfaces inspection and it's benefits compared to the handcrafted inspection, we will explain the wavelet transformation method,with an emphasis on it. There are various methods for image segmentation. Yet, in this essay, we will use the wavelet transformation for segmentation of steel colorful plates to areas of normal and defective. Each image needs to be converted to RGB, HSL, LAB color spaces. Afterward, considering a color space, discrete wavelet transformation is applied to three dimensional channels of the image and detailed images at various levels are obtained. Due to visible differences in normal and defective areas, it is expected that defective areas have clear borders with normal areas in some detailed images and in some way clustering the image to the areas of defective and normal be possible. Finally, the results obtained from different colorful channels are compared. It is worth mentioning that the tests have been done on a set of images of normal and defective steel surfaces, showing the quality of wavelet method.**

Keywords-Color; Image segmentation; Metals industry; Defect detection; Wavelet Transform; Steel Surfaces Inspection; color spaces;

I. INTRODUCTION

Quality control of products has always been of importance in the highly competitive stainless steel industry. The users of stainless steel set ever-increasing requirements on product quality. Many material properties can still only be measured in laboratory, however more and more measurements are now made on-line throughout the production line. Impartially especially surface defects have to be detected on-line with a surface inspection system because of their random appearance.

A number of methods of combining color and texture descriptions have been proposed [9, 10]. These and many other methods treat color and texture as a joint phenomenon. The purpose of a surface inspection system is to detect and classify surface defects that impair product quality regarding the standards and requirements set by the user. The requirements mostly deal with the suitability of the product to the intended use of it. In the worst case the defects may make the product functionally deficient or even unusable. Critical defects are also those which cause production disturbances. There are many potential areas of application for texture analysis in industry [1-4], but only a limited number of examples of successful implementation of texture in inspection exist. These systems utilized various techniques for defect detection. At best their sensitivity was good. Steel surface producers would like to know emergence of such defects that may prevent running the process smoothly and to make sure that the product quality meets the customer requirements. The causes of different defects should be located and removed as soon as possible. If defects are present, they also should be recorded for different types of statistical quality reports. Without a surface inspection system, surface defect identification and root cause tracing can easily take a long time. During this delay the problem may repeat causing even more downgraded production. Surface inspection and quality classification of steel is an essential stage in the steel manufacturing industry. Due to the high cost of human inspection, speed of the production line, and repetitious nature of the activity, development of an automatic inspection and defect detection system would have an impressive impact on the overall performance

of a steel production plant. Typically, defects (e.g. a black patch, hole, indent) affect the expected texture of the steel and hence can signify a 'textural abnormality'. Fig. 1. shows four samples of normal and defective steel surfaces.

Traditionally, many annealing and pickling lines have visual inspectors to do the surface inspection. Compared to this method, automatic surface inspection offers several benefits. As a summary, important facts provided by a steel inspection system are:

- Providing consistent high performance surface quality inspection.
-Reducing costly reject production and customer claims
- Improving real-time process control
- Identifying preventive maintenance
- Facilitating quality grading of products
- Defect statistics for quality follow-up

II. BACKGROUND

The starting point for the segmentation procedure is the image captured by the camera or scanner. Mostly surface inspection systems are used directly from digital cameras. Figure 2 shows the overall process scheme of an image surface of steel parts using wavelet transform approach. The block is clear that the first image is selected from the dataset and after necessary preprocessing (noise removal),the image is converted to the desired hsl, lab color space. Then discrete wavelet transform is separately applied on the images in different color spaces. Edges in Images defective areas obtained from wavelet transformion marked in the image And defective areas are clustering And are different from adjacent pixels. After clustering is to obtain images in three color space, in order to identify defective areas on the original image so that effect with the test method is visible to the eye Or the percentage of correct detection and faulty healthy areas to be done properly, it is necessary operations post- processing done. In this operation, the defective areas specified boundary, areas to separate the two parts are divided into normal and defective. This image to determine the percentage of defective parts and to determine percentage of defective parts on the original image is used. Texture analysis techniques such as the occurrence matrices or frequency domain methods, due to the complexity of computing time, systems are not suitable for real-time

inspection. Method used in this study in terms of time on the above-mentioned methods has advantages. [5,14,15,16].

III. PROPOSED METHOD

Discrete wavelet transform

There are various methods for image segmentation that we have here discrete wavelet transform method for steel-colored pages segmentation have been using. Variety of methods to detect defects and tissue classification images based on wavelet transform are used. [6,7,8]. Our three color channels RGB and HSL and LAB separately and the results applied wavelet edge detection and ultimately will offer segmentation. in wavelet transform method and a defect detection, there is no special pre processing. Because the causes of defects or areas of light and healthy with these changes are known as the defective areas. But the noise in the images should be deleted where here salt and pepper noise with the median filter will remove the image noise. Discrete wavelet transform was sufficiently accurate and the indiscriminate increase in volume of information and reduces computation time. Wavelet transform applied to two sets of high pass and low pass filter on the image. Passing image of these filters for the first time, four new image is obtained: Approximation (low pass filter result), horizontal detail in the points, points of vertical detail, and detailed in diameter (high pass filter results). In order to calculate the wavelet coefficients in the two layers, for the second time approximation image obtained from previous stage must be passed through filters. This process can be repeated again. Each time employing wavelet, is half the size of the resulting images. Wavelet transform of an image, change brightness scale is much different. For example, wavelet coefficients in the structural image edges, in areas with the maximum value is defective[11]. Therefore, applying wavelet transform on the desired two-dimensional image, points that have fractures or severe scratches and any fluctuation in the tissue is a white line (maximum light intensity) in the segmentation will have the final image.
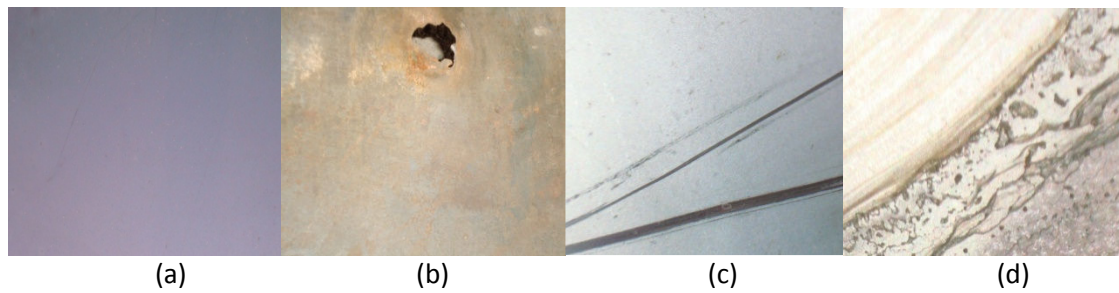
| (a) | (b) | (c) | (d) |

Fig1.   (a) Normal surface (b) hole abnormal (c) scratch abnormal and (d) black patch abnormal

While color images segmentation, fracture points and edges, clustering and a distinct color from the color pixels are nearby. We know that wavelet, assessment of similarity between the frequency content (scale) signal and wavelet function at different scalesIn case of discrete wavelet transform, filters with different cut off frequencies for different scales of signal analysis is used. As a result of crossing signals and low pass filters, high pass, different frequencies are analyzed. Had to be careful that the number of steps needed for the discrete wavelet, depend on the the analyzed signal frequency characteristics. Finally, the discrete wavelet signal combination to filter output from the first stage of filtering actions are obtained. Thus, the number of wavelet coefficients with the number of discrete input signal samples will be equal. To apply wavelet transform on the images, should we use two-dimensional wavelet transform. For this purpose apply the transform wavelet of one-dimens ional matrix of rows and columns, so we apply a combination of these two components convert, convert two-dimensional achieved. The process in Figure 3 has been shown[12].

In this figure, the initial image pass along x (rows) of a low-pass filter and a high pass filter and be sampled decreases. This phase will result in two files. One image contains low frequency IL (x, y) and the other including IH (x, y) image is high frequencies. In the next stage, each of these two image pass along y (columns) from a low pass filter and a high pass filter and be sampled decreases. Thus the following four image can be produced:

  - ILL quantity corresponds with the quantity of low-frequency files in both direction.

  - ILH quantity is includes of the image horizontal details.

  - IHL quantity is includes of the image vertical details.

  - IHH quantity also included details of diameter.

Studies showed that Chang and Kuo [13] Much of that information in the areas of tissue are intermediate frequency. Therefore, proper texture analysis is that detailed images from the wavelet transform in such a manner that we choose is mostly comprised of middle frequencies.

IV- DATASET AND EXPERIMENT

Performance with our proposed method applied on the steel surface images will look. Images obtained are four collections, are including images of healthy and defective (including defect:hole - black patch- scratch) The collection includes images of handmade images, which specific defects in their creation have put to the test. Our input images are in RGB format and all with relatively high resolution have been prepared. Images obtained are converted to images smaller in size 512 * 512 pixels. Total images taken from steel surfaces, 13 files have been tested. Images are selected for testing involved a variety of defects. To convert these images to LAB and HSL spaces we use relations are listed in [14].
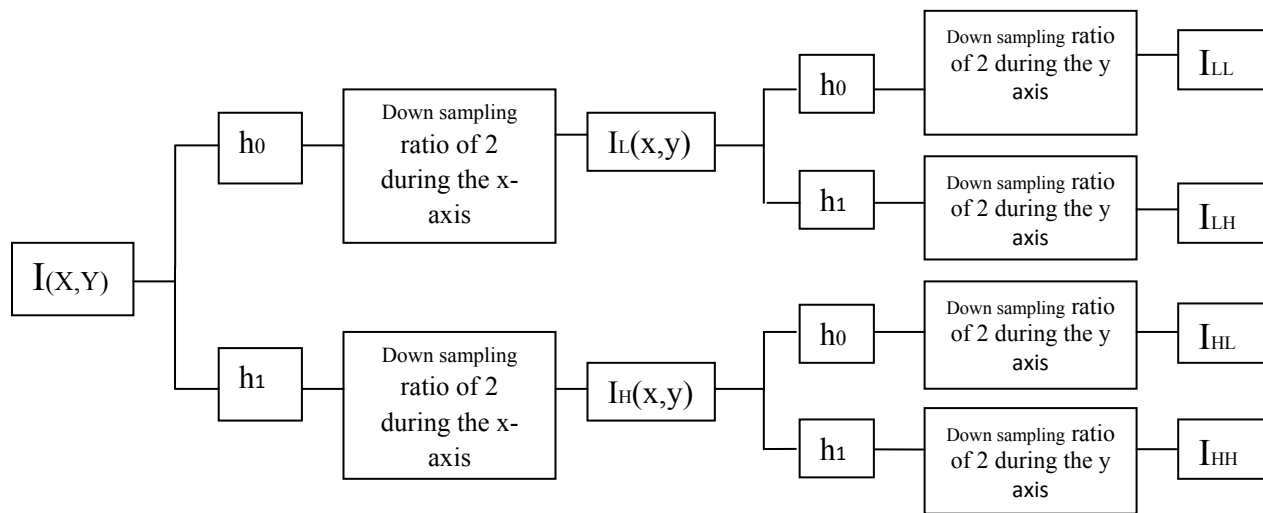
Figure (3). Block diagram of an image wavelet transform in two scales

A. handmade steel Images segmentation with specific defects

First hand pictures were used to test the method by applying wavelet transform on the images to segmentation images with specific defects in three color models, RGB HSL, LAB have paid. Table (1), including amounts that SNS and SPC, respectively, show percentage correct detection and the percentage of defective area detection. which SNS,is called sensitivity and SPC is called Specifity.

[1]SNS : Percent to detect defective areas total image area

[2]SPC: Percent to detect Correct areas total image area

$$SNS - TPP - \frac{TP}{TP+FN} - \frac{TP}{P} \quad (1)$$

$$SPC = \frac{TN}{FP+TN} = \frac{TN}{N} = 1 - FPR \quad (2)$$

The above criteria can be used to calculate efficacy. P indicates that the desired class and N class is undesirable.

• If the samples belong to positive class and the method will be in the positive class, this sample, a positive sample has been correctly guessed (TP).
• If the samples belong to negative class and method will be in the negative class, in this instance, an example of a false negative has been conjectured (FN).
• If the samples belong to negative class and the method will be negative class, this sample, a negative sample is guessed correctly(TN).
• If the samples belong to negative class and method will be negative class, this sample, an sample of a false positive has been conjectured(FP).

---

[1] Sensivity
[2] Specifity

TABLE I. THE RESULTS FOR THE TEST OF MANUALLY IMAGE

Image no	SNS			SPC		
	RGB	HSL	LAB	RGB	HSL	LAB
Image1	%100	%99	%98	%99	%98	%99
Image2	%100	%98	%97	%98	%99	%98
Image3	%100	%96	%93	%98	%98	%99
Image4	%98	%95	%93	%95	%97	%98
Average	%99	%97	%95	%96	%98	%98

In this test saw that the correct diagnosis of defective pixel RGB color model better than the HSL model and LAB has done. Also identify areas safe for the percentage of LAB color model is better than other models. Such images due to defects identified manually and better results than the images will have a real defect. Collection of images with hand defects for review and evaluation procedures done properly used. Figure (4) results of visual approach on a sample defect images with hand colored in three models shows.

B. *steel image segmentation with hole defection*

How to do this test first test is quite similar with the difference that a series of images with a real defect defect cavity instead of handmade pictures have been using. Each of these images Judicial RGB, HSL, LAB tested separately have. Table (2) Test results related to the specific defect cavity with images in three color channels shows.

TABLE II. TESTING RESULTS RELATED TO HOLE DEFECTIONS

Image no	SNS			SPC		
	RGB	HSL	LAB	RGB	HSL	LAB
Image1	%96	%94	%90	%92	%98	%98
Image2	%97	%92	%87	%88	%91	%97
Image3	%95	%91	%97	%89	%91	%95
Average	%96	%92.3	%91.3	%89.6	%93.3	%96.6

Considering the results is evident that performance test as well as the first test. Because the defect detection in RGB color channel is better than HSL and LAB. Also, safe areas detection in LAB color channel is better than Channel HSL and RGB. Difference this test with previous test in this image is being real.

C. *steel image segmentation with black patch defection*

In this experiment the black patch images (spot the non-normal) will examine. Most existing methods in the detection of black patch defects are poor, But this method to detect this type of defect, better than anyone else has done methods and has a higher accuracy than other methods. Finally calculate the percentage of SNS and SPC to review the results explains. Calculated trend in all the tests done matlab software to be tested can be compared fairly different. RGB space in this test in diagnosis of defects of steel surfaces is better than other extinction.

TABLE III. RESULTS FOR TEST IMAGES WITH BLACK PATCH DEFECTION

Image no	SNS			SPC		
	RGB	HSL	LAB	RGB	HSL	LAB
Image1	%95	%89	%86	%97	%96	%98
Image2	%97	%88	%83	%98	%97	%97
Image3	%97	%91	%84	%96	%98	%94
Average	%96.3	%89.3	%84.33	%97	%97	%96.3

D. *steel image segmentation with scratch defection*

Last category, images are scratched. This type of defect the most common type is a steel surface defects. What a perfect expression of this test is all surface scratches and elongation that there are in effect Inappropriate produce or surface contact.

| (4) | (3) | (2) | (1) |
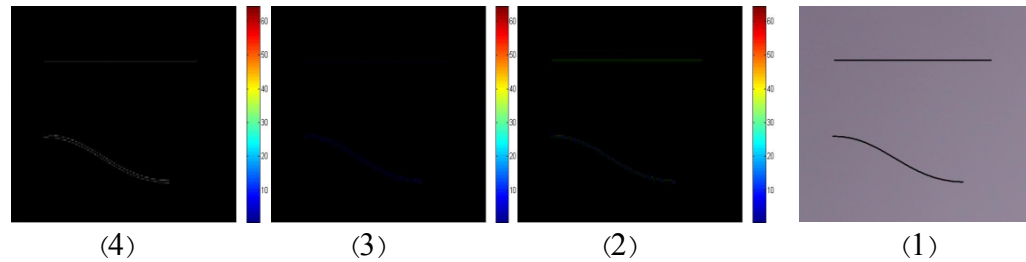
Figure (4): results related to the method applied on a handmade picture with specific defects in three color channels (1) original (2) image segmentation in RGB space (3) image segmentation in HSL space (4) image segmentation in the LAB space.

TABLE IV. TESTING RESULTS TO DETERMINE SCRATCH DEFECTION

Image no	SNS			SPC		
	RGB	HSL	LAB	RGB	HSL	LAB
Image1	%96	%92	81.96	%94	%98	%98
Image2	%92	%84	%47	%95	%96	%97
Image3	%93	%88	%89	%94	%96	%95
Average	%93.6	%88	%72.6	%94.3	%96.6	%96.8

Table (4) is shown The results of defect detection of scratch images using wavelet transform. Percentage of SNS and SPC for images in RGB space is better than HSL space and HSL space is better than LAB space. The percentage of SPC in this test for LAB space is better than HSL space and HSL space is better than RGB space. Overall, we can say in this test, like before tests, defect detection in RGB space is better than other places. In the table below the overall defect detection images in three color channels with the proposed method is presented.

Table (5) shows Overall results and the average from the previous tests. In general, the above table, we find that the correct diagnosis of defective areas in a RGB space is better than HSL color space. The detection of correct areas in HSL space is better than RGB space. RGB space in the diagnosis, Is better than HSL, LAB. that HSL color space is better than LAB color space. Therefore, the sum of these results can be the best color space in the steel surface defect detection is RGB space. Because this space in all experiments is shown the best detection results.

TABLE V. RESULTS FOR ALL IMAGES TO DETECT DEFECTIONS IN THREE DIFFRENT COLOR CHANNELS.

Color Space	SNS			SPC		
	RGB	HSL	LAB	RGB	HSL	LAB
Average	%95.3	%89.8	%81.95	%93.6	%95.6	%96.5

A   B   C   D
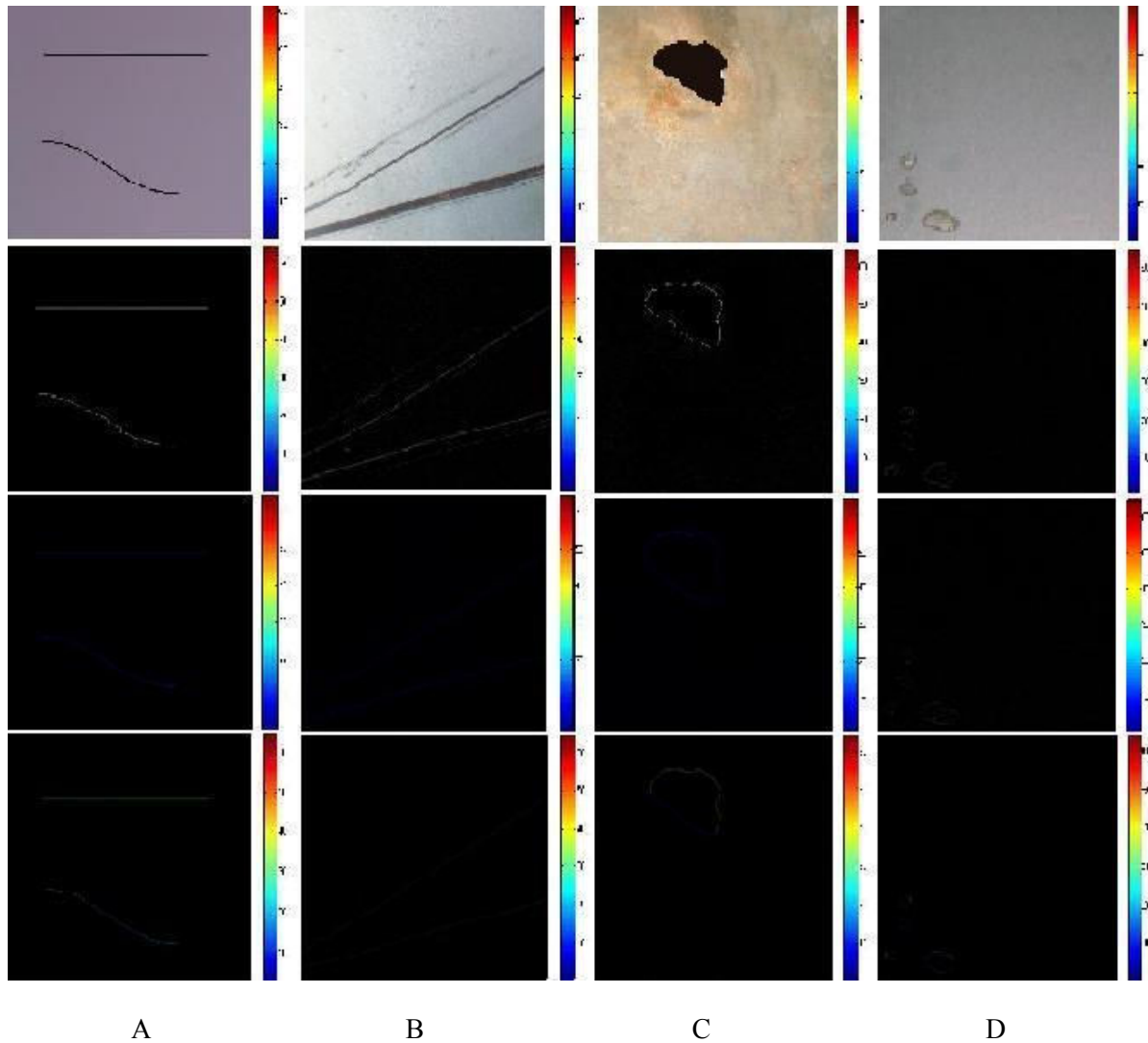
Figure (5) results from applying the proposed method on images with various defects in three color channels RGB, HSL and LAB. first row is Original image- second row is segmentation in the RGB space –third row is segmentation in HSL space - fourth row is segmentation in LAB color space. Column (A) includes handmade image segmentation - Column (B) contains scratch image segmentation - Column (C) contains hole image categories - Column (D) includes image segmentation with black spatch.

Figure (5) is shown experiments with the proposed method. The image in column (A) includes an example and handmade images are segmented with existing methods in RGB and HSL and LAB channels. Images in (B) Column includes scratch defect images that results segmentation are visible. Similarly, the (C) and (D) column images show hole and black patch defect. Efficiency of the proposed method of images is diagnosed.

V.CONCLUTION

In this paper, The wavelet transformation method was used to find the image edge as to detect the normal and defective areas of steel surfaces. In This method, It is not required to have previous information about the images nor to system education for defects diagnosis. It can be said that, unlike segmentation methods acting on pixels, the wavelet transformation method, tested on three colorful models in this paper, acts on the whole image, similar to the human eyes. Of advantages of wavelet transformation method in comparison with other methods, one can refer the relative speed increase in this method. This proposed method is without any supervision and has the privilege of diagnosing the defective and normal areas with a high percentage. Besides, in terms of visibility, this method performs the segmentation of different images with high accuracy. Unlike most other

methods which can only detect specific types of defects, the output image is segmented and the results of defects detection is shown for all images and all defects with a high percentage of accuracy. As observed, color plays a great part in defect detection. In the obtained results, the RGB color space was observed to be the best for detection of steel surfaces faults.

REFERENCES

[1]   T.S.Newman and A.K.Janin, A survey of automated visual inspection, comput. Vision Image Understanding 61(1995)231-262.

[2]   M.Pietikainen and T.Ojala, Texture analysis in industrial applications, in Image Technology-Advances in Image Processing, Multimedia and Machine Vision, J.L.C.Sanz,(1996) 337-359.

[3]   K.Y.Song,M.Petrou and J.Kittler, Texture defect detection: areview, SPIE vol.1708 Application of Artificial intelligence X:Machine Vision and Robotics, (1992),99-106.

[4]   M.Pietikainen et al, Approaches to texture-based classification, segmentation and surface inspection, in Handbook of Pattern Recognition and Cimputer Vision, 2nd edition, eds. C.H.Chen,L.F.Pau,P.S.P.Wang,(1999) 711-736.

[5]   R.Kruger, W. Thompson, and A.Turner. Computer diagnosis of neumoconiosis. IEEE Transactions on Systems, Man, and Cybernetics,4(1):40-49,1974.

[6]   H. Y. T. Ngan, G. K. H. Pang, S. P. Yung, and M. K. Ng, Wavelet based methods on patterned fabric defect detection, Pattern Recognition, vol. 38, pp. 559-576, Apr 2005.

[7]   N. Sebe, M.S. Lew, Wavelet based texture classification, Pattern Recognition, Proceedings 15th International Conference on ,Vol. 3, Page(s):947 – 950 ,2000.

[8]   S. Arivazhagan, L. Ganesan, V. Angayarkanni, Color texture classification using wavelet transform, Computational Intelligence and Multimedia Applications, Sixth International Conference on, 16-18 Aug. 2005 , Page(s): 315 – 320, 2005.

[9]   Jain A and Healey G. A multiscale representation including opponent color feature for texture recognition IEEE Transaction on Image Processing Jain. 1998. 7(1):124128.

[10]   Rimac-Drlje, A. Keller, Z. Hocenski,, Neural Network Based Detection of Defects in Texture Surfaces, Proceedings of the IEEE International Symposium on Industrial Electronics, Vol. 3, Page(s): 1255 – 1260, June2005.

[11]   Mirmehdi M and Petrou M. Segmentation of color textures. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000:22(2):142-159.

[12]   Sidney Burrus C, Gopinath RA, Introduction to wavelets & wavelet transforms, Prentice Hall,   New Jersey, 1998.

[13]   Chang T, Kuo J, Texture analysis & classification with tree-Structured wavelet transform, IEEE Trans. Image Processing, 1993; 2(4) : 429-441.

[14]   A.Monadjemi, Towards efficient Texture Classification and Abnormality Detection, Phd thesis of Univiersity of Bristol , October 2004

[15]   Kamal Jamshidi, S Amirhassan Monadjemi, and hesam hasanpour. Steel Surface Inspection Using Local Binary Pattern and Color Features. *ICCTA2006*, 1384

[16]   A.Monadjemi, J.Alemasom. Steel defect detection using New Gabor Method,5th image processing & mashin vision conf. Iran, MVIP 2008.

AUTHORS PROFILE

Seyed Amirhassan Monadjemi, born 1968, in Isfahan, Iran. He got his PhD in computer engineering, pattern recognition and image processing, from University of Bristol, Bristol, England, in 2004. He is now working as a lecturer at the Department of Computer, University of Isfahan, Iran. His research interests include pattern recognition, image processing, human/machine analogy, and physical detection and elimination of viruses.

Clustering in Mobile Ad hoc Networks: A Review

Meenu Chawla

Department of CSE
MANIT, Bhopal, India
chawlam@manit.ac.in
meenu_chawla_manit@rediff.com

Jyoti Singhai

Department of ECE
MANIT, Bhopal, India
j_singhai@manit.ac.in

J L Rana

Department of CSE
MANIT, Bhopal, India
ranajl@manit.ac.in

Abstract—**Mobile Ad-hoc Networks (MANETs) are future wireless networks consisting entirely of mobile nodes that communicate on-the-move without base stations. Nodes in these networks generate user and application traffic and carry out network control and routing functions. Dynamic and random topologies lead to rapidly changing connectivity and network partitions. This dynamic nature along with bandwidth and power constraints together pose new problems in network scalability, network control, especially in the design of higher level protocols such as routing, and in implementing applications with Quality of Service requirements. Hierarchical routing provides a means to tackle the above mentioned problems in large scale networks. Clustering is the process of building hierarchies among nodes in the network. In this approach an ad hoc network is partitioned into group of nodes called as clusters. This paper presents a review of the different clustering algorithms and the criterion on the basis of which each of them takes the clustering decisions.**

Keywords- Mobile Ad-hoc networks; clustering; clusterhead selection.

I. INTRODUCTION

In hierarchical routing the nodes in the network are dynamically organized into partitions called clusters, and then the clusters are aggregated again into larger partitions called super clusters and so on. The nodes geographically close to each other form a cluster. Each cluster elects a leading node called the cluster head which acts as a coordinator for the cluster. The nodes connected to more than one cluster are called gateway nodes and act as relays between clusters. Dividing a network into clusters helps maintain a relatively stable network topology. Clustering makes network more manageable. Cluster size is controlled through the radio transmission power.

Cluster based algorithms are among the most effective routing algorithms due to their scalability[1,2,26]. Clustering outperforms other routing algorithms in case of large networks. As all inter-cluster routing in such a scenario is through the cluster head, it is therefore more burdened than its members and tends to be a bottleneck in the system if not chosen appropriately. The objective of any clustering algorithm is to partition the network into several clusters which is the focus of current literature in this area.

Several algorithms have been suggested for clustering and clusterhead selection. A number of clustering algorithms have been proposed, some very simple[3,4,5] and some with a view

of optimally utilizing the critical parameters[6,7,8,9,10,14] of ad hoc networks. A review of the clustering and cluster head selection algorithms is being done in this paper.

II. REVIEW OF CLUSTERING ALGORITHM

A. Lowest ID algorithm

The Lowest-ID algorithm [3, 16] is the simplest clustering algorithm .In this algorithm every node in the network has a unique identifier (ID). Nodes periodically broadcast their ID in "hello messages". Each node compares the IDs of its neighbors with its own ID, than a node having lowest ID decides to become a cluster head. The algorithm takes following steps:

1. Every node broadcast its own ID periodically in Hello message.

2. All nodes receive hello messages from their neighboring nodes and match their IDs then the node having lowest ID is elected as cluster head.

3. The node, which can hear broadcast from two cluster head, is than becomes gateway node.

In this algorithm there is no limit to the member nodes that a cluster can have. No network related parameter is given any consideration in selection of clusterhead, and hence the performance of such networks is of random and unpredictable nature.

B. Highest Degree Algorithm

The Highest-Degree heuristic [3,4] takes into account the degree of a node, i.e. the number of its one-hop neighbors. Each node periodically broadcasts its degree value. A node with the highest value of degree in its neighborhood is selected as the cluster head and its neighbors join it as cluster members. The procedure is repeated with the remaining nodes until each node is assigned to a cluster. Any tie is broken by the lowest id criterion. This heuristic doesn't put any upper bound on the number of nodes in a cluster, consequently the cluster head becomes highly overloaded leading to performance degradation. As the network topology changes, this approach can result in a high turnover of cluster heads. This is because when the highest connectivity node drops even one link due to node movement, it may fail to be re-elected as a cluster head.

These are two most popular criteria to partition mobile nodes. Both these algorithms do not provide any quantitative measure of cluster stability.

Corresponding author: Meenu Chawla

C. Least Cluster head Change Algorithm

k-CONID [5] combines the two approaches Highest Degree and LowestID. Connectivity is considered as a primary and lower ID as a secondary criterion for selecting cluster heads. The algorithm considers at most k hop neighbours of a node for cluster head selection. At the beginning of the algorithm, a node starts a flooding process in which a clustering request is send to all other nodes. In the Highest-degree heuristic, node degree only measures connectivity for 1-hop nieghbours. k-CONID generalizes connectivity for a k-hop neighborhood. Thus, when k = 1 connectivity is the same as node degree.

Each node in the network is assigned a pair: dID = (d, ID). d is a node's connectivity and ID is the node's identifier. A node is selected as a cluster head if it has the highest connectivity. In case of equal connectivity, a node has cluster head priority if it has lowest ID. Every node broadcasts its clustering decision only after all its k-hop neighbors with higher value of (degree, id) pair have broadcast their clustering decision.

Although each node determines one cluster, clusters may overlap. This means that a node can belong to all clusters whose cluster head is at most k-hops distance from the node. Nodes that belong to more than one cluster become gateway nodes.

D. (α,t) Cluster Framework

McDonald and Znati[6] have proposed a framework for dynamically organizing mobile nodes in a MANET into clusters which has been called the (α,t)-cluster framework. The approach is to maintain topology which allows for optimal routing in face of low mobility and efficient routing if node mobility is high. Here the focus is on mathematical characterization of the probability of link and path availability as a function of a random walk based mobility model [21]. In the (α , t) approach it is attempted to provide an effective topology that adapts to node mobility.

Path availability is a random process which is determined by the mobility of nodes that lie along a certain path. In the (α , t) approach paths are evaluated by two system parameters, α and t. α establishes a lower bound on the probability that a given cluster path will remain available for a time t. α controls cluster stability while the role of t is to manage cluster size for a given level of stability.

The actions taken by the clustering algorithm depend upon the information given by the routing and network-interface layer protocol. Each node in the network is given a node's cluster identifier number (CID) and makes use of a timer named α timer. This timer establishes the maximum time t for which a node guarantees that paths will be available to each cluster destination with probability = α.

In the (α, t) algorithm, clusters which satisfy the (α, t) criteria are maintained. The (α, t) criteria is accomplished if the probabilistic bound α on the mutual availability of paths between nodes in a cluster exists over a specified interval of time t. Therefore, the algorithm applies prediction of node mobility as criteria for cluster organization. The (α, t) algorithm

characterizes the probability of link and path availability as a function of a random walk mobility model.

The algorithm is designed to take appropriate actions upon topological changes. A topological change requires that nodes revaluate the (α, t) criteria. The documentation that supports this clustering approach presents the pseudo code for five important topological changes that determine the (α, t) cluster algorithm: Node activation, link activation, link failure, node deactivation and α timer expiration.

It has been shown that the (α,t)-cluster strategy has been effective in terms of adapting to node mobility, achieving node stability in face of mobility and protocol efficiency.

E. MOBIC

Basu et.al [10] proposed a variant of Lowest-ID algorithm, MOBIC, which is similar in execution to the Lowest-ID algorithm except that the mobility metric is used as a basis of cluster formation instead of ID. MOBIC uses a new mobility metric; Aggregate Local Mobility (ALM) to elect CH. ALM is computed as the ratio of received power levels of successive transmissions by transmitting periodic 'hello' messages, between a pair of nodes. This gives a measure of relative mobility between neighbouring nodes. Each node then calculates aggregate local mobility metric M value by calculating the variance (with respect to zero) of the entire set of relative mobility samples of all its neighbours. The node with lowest value of M becomes clusterhead.

The main drawback of this algorithm is that it uses signal strength as a measure of node mobility. However, because of noise, obstacles, variation in battery power, etc, weight based on variation in signal strength may not be accurate, so stability of a node can not be evaluated clearly. Although mobility is one of the most important factors that can affect the stability of a clusterhead, there are other equally critical parameters that need to be considered for stable clusterhead selection. Considering only a single parameter will not give desired stability in all types of scenarios. Also the algorithm here is just looking at the stability of the clusterhead alone, and not at the stability of the complete network. To ensure stability of the entire network, consideration for stability of gateway nodes is of importance

F. MobDhop:

A distributed clustering algorithm called MobDhop[9] has been reported which partitions an ad hoc network into d-hop clusters based on a mobility metric. The objective of forming d-hop clusters is to make the cluster diameter more flexible. MobDhop is based on mobility metric and the diameter of a cluster is adaptable with respect to node mobility. This clustering algorithm assumes that each node can measure its received signal strength. In this manner, a node can determine the closeness of its neighbors. Strong received signal strength implies closeness between two nodes. The MobDhop algorithm requires the calculation of five terms: the estimated distance between nodes, the relative mobility between nodes, the variation of estimated distance over time, the local stability, and the estimated mean distance. A node calculates its

estimated distance to a neighbor based on the measured received signal strength from that neighbor. Relative mobility corresponds to the difference of the estimated distance of one node with respect to another, at two successive time moments. This parameter indicates if two nodes move away from each other or if they become closer.

The variation of estimated distances between two nodes is computed instead of calculating physical distance between two nodes. This is because physical distance between two nodes is not a precise measure of closeness. For instance, if a node runs out of energy it will transmit packets at low power acting as a distanced node from its physically close neighbor. The variation of estimated distance and the relative mobility between nodes are used to calculate the local stability. Local stability is computed in order to select some nodes as cluster heads. A node may become a cluster head if it is found to be the most stable node among its neighborhood. Thus, the cluster head will be the node with the lowest value of local stability among its neighbors.

MobDhop is executed in three stages as follows:

- Discovery stage: At the initialization of the network, two hop clusters are to be formed in this stage. For this the nodes exchange hello messages periodically which includes the local stability value of the node(initialized ti infinity). After a discovery period in which nodes acquire complete knowledge of their neighbour nodes, each node computes its local stability value and broadcasts it for information to its neighbours. Node with lowest value of local stability becomes cluster head and its local stability value is the group stability (GS). If a node can hear messages from a node that belongs to a different cluster, it becomes a gateway node. If not, it becomes a cluster member.

- Merging stage: The two-hop clusters established in the discovery stage are expanded by a merging process. A merging process can be initiated by a nonclustered node that requests to join its neighbouring clusters or when two neighbouring gateways request to merge their clusters. The merging is allowed only if the two merging criterion as stated are fulfilled. First condition ensures that the variation of estimated distance between two merging nodes should be less than or equal to the minimum value of group stability of the two clusters. Second condition states that the mean distance between two gateways should be less than or equal to the higher value of estimated mean distance of the two clusters. This is to ensure that the distance characteristics of the clusters are met.

- Cluster maintenance stage: A cluster maintenance stage is invoked when topology changes occur due to either arrival of a new node or a node leaving the network. When a node switches on it will begin the merging process as described in order to join a cluster. When a node which is a clusterhead switches its immediate neighbours begin the discovery process as described so that a new cluster head can be selected. During the period when the nodes are without a clusterhead (clusterhead election period) the two hop neighbour nodes initiate a merging process and join other clusterheads if the merging criterion is met.

This algorithm also suffers from the same drawback as MOBIC. The algorithm here is just looking at the stability of the clusterhead alone, and not at the stability of the complete network. To ensure stability of the entire network, consideration for stability of gateway nodes is of equal importance. If gateway nodes are highly mobile then intercluster routes will break frequently leading to frequent re-routing causing high routing overhead. Also only the mobility criterion is taken into account for determining the stability of the network. Other parameter such as remaining battery power of the node is an important parameter which affects the stability of the network and should be considered.

G. DMAC

The Distributed and Mobility-Adaptive Clustering (DMAC) [7] algorithm provides a generalized solution for clustering framework. Nodes are assigned weights based on nodes' mobility-related parameters. The weights express how suitable a node is for the role of cluster head given its own current status. The bigger a node's weight, the more suitable it is for the role of cluster head. This implies that, when due to the mobility of the nodes two or more cluster heads become neighbors, those with the smaller weights have to resign and affiliate with the now bigger neighboring cluster head. DMAC overcomes a major drawback found in most clustering algorithms. A common assumption that is presented in most algorithms is that during the set up time nodes do not move while they are being grouped into clusters. Normally, clustering algorithms partition the network into clusters and only after this step has been accomplished, the non mobility assumption is released. Afterwards, the algorithm tries to maintain the cluster topology as nodes move. In real ad hoc situations this assumption can not be made due to the constant mobility of nodes. Therefore one important feature of DMAC is that nodes can move, even during the clustering set up.

During the algorithm execution it is assumed that each node has a weight (a real number $= 0$) and an ID (node's unique identifier) associated to it. The weight of a node represents node mobility parameters. A node chooses its own role (cluster head or ordinary node) based on the knowledge of its current one hop neighbors. A node becomes a cluster head if it has the highest weight among its one-hop neighbours; otherwise it joins a neighbouring cluster head.

During execution of this algorithm, every node has knowledge of its ID, its weight as well as its neighbors ID and its neighbor's weight. DMAC is a message driven algorithm (except during the initial procedure). Two types of messages are used: If a node joins a cluster it sends out a Join message and if it becomes a cluster head it sends a CH message. A node decides its own role once all its neighbors with bigger weights have decided their roles.

DMAC executes five procedures at each node: an init routine, a link failure procedure, a new link procedure, a procedure depending upon the reception of a CH message and a procedure depending upon the reception of a Join message.

When a cluster head receives a Join message from an ordinary node, it checks if the sending node is joining its own cluster or a different one. On the other hand, if an ordinary node receives a Join message from its cluster head, it means that this cluster head is giving up his role. Upon the reception of a CH message a node checks if it will affiliate or not to the sending cluster head.

The adaptation feature of the clustering algorithm is made possible by letting each node react to the failure of a link with another node or to the presence of a new link. Upon link failure between a cluster head and one of his node members, the membership of the node to the cluster is removed, and this node must determine its new role. A new link between two nodes means that a node has detected the presence of a new neighbour. In this case, the node must determine if this new node has a larger weight than its own cluster head in order to join it. If the node is a cluster head itself then it will give up its role if the new cluster head has a higher weight.

Although DMAC provides a generalized framework for clustering nodes, it does not specify clearly weight metric method.

H. A Weighted Clustering Algorithm (WCA)

Although DMAC provides a generalized framework for clustering nodes, it does not specify clearly weight metric method. A distributed clustering algorithm based on weight values has been proposed by M. Chatterjee et.el.[8][24]. The weighted clustering algorithm (WCA) selects clusterheads, according to the number of nodes it can handle, mobility, transmission power and battery power. To avoid communications overhead, this algorithm is not periodic and the clusterhead election procedure is only invoked based on node mobility and when the current dominant set is incapable to cover all the nodes. To ensure that clusterheads will not be over-loaded a pre-defined threshold is established in order to specify the number of nodes each clusterhead can ideally support. This parameter corresponds to d. WCA selects the clusterheads according to the weight value of each node. The weight associated to a node v is defined as:

The node with the minimum weight is selected as a clusterhead. The weighting factors are chosen so that $w_1 + w_2 + w_3 + w_4 = 1$. M_v is the measure of mobility. A node with less mobility is always a better choice for a clusterhead. Here mobility is taken by computing the running average speed of every node during a specified time T. Δv is the degree difference that depicts the deviation of the actual degree of a node to the proposed ideal degree. D_v is defined as the sum of distances from a given node to all its neighbours and is used to depict energy consumption parameter since more power is needed for larger distance communications. The parameter P_v is the total time the node has served as a clusterhead. P_v is used to give a measure of how much battery power has been consumed. Power consumption in a clusterhead is higher than in an ordinary node as is has extra responsibilities. The total weight here is calculated such that it gives a deviation value from the ideal conditions hence the node with minimum weight is elected as cluster head. An attempt to optimize WCA by using entropy [10] has been made.

1) Optimization of WCA using Genetic Algorithm:

In [11], genetic algorithms have been used for enhancing the performance of clustering algorithms in mobile ad hoc networks. The clustering problem is mapped to individual chromosomes as input to the genetic algorithmic technique. The authors have particularly optimized the popular WCA algorithm. In the genetic algorithm used each chromosome contains information about the cluster heads and the members thereof, as obtained from the original WCA. The genetic algorithm then uses this information to obtain the best solution (chromosome) defined by the fitness function. The proposed technique is such that each cluster head handles the maximum possible number of mobile nodes in its cluster in order to facilitate the optimal operation of the medium access control (MAC) protocol. Consequently, it results in the minimum number of clusters and hence cluster heads. Simulation results exhibit improved performance of the optimized WCA than the original WCA.

2) Optimization of WCA using Simulated Annealing:

In [12], simulated annealing algorithm has been applied to clustering algorithms used in ad hoc networks. Determining the optimal dominant set is a NP-hard problem. A good polynomial time approximation algorithm may be used to obtain a near optimal dominant set for cluster based topology. Simulated annealing is a probabilistic search technique that has been used to solve a large number of combinatorial optimization problems in engineering and science. In [27] Kirkpatrick et. al, has applied simulated annealing to two problems: the physical design of computers and traveling salesperson problem, both of which are NP problems. He says that any new combinatorial optimization problem can be solved easily using simulated annealing by specifying four main factors of the algorithm: a concise description of a configuration of the system; a random generator of "moves" or rearrangements of the elements in a configuration; a quantitative objective function containing the trade-offs that have to be made; and an annealing schedule of the temperatures and length of times for which the system is to be evolved.

The initial solution given by original WCA has been mapped to simulated annealing algorithm in order to find the best possible solution from a set of all possible set of dominant sets. If a node is neither a cluster head nor a member of any cluster and its degree is less than a threshold value than the node is added in the random dominant ser. The objective function is chosen to be the sum of weights of nodes in the dominant set. The fitness value is computed in order to obtain the best solution by comparing each solution to the current best. This algorithm eventually finds the best solution without getting stuck in local minima by following the simulated annealing algorithm. Hence simulated annealing optimizes the performance of WCA such that each cluster head handles the maximum possible number of nodes in its cluster, resulting in minimum number of cluster heads and clusters. The simulation results show that fewer cluster heads are obtained by applying simulated annealing to WCA than the results of the original WCA.

I. *Efficient Management Algorithm for Clustering (EMAC)*

EMAC [20] is another distributed clustering algorithm where the node weight is calculated using factors like the node degree, remaining battery power, transmission power, and node mobility for the cluster heads' election. The algorithm uses transmission range of node Pi instead of the sum of distance used in WCA in order to elect the node which can cover the largest range, thus, minimizing the number of clusters generated. In addition, the author argues that remaining battery power is a better measure than the cumulative time during which the node acts as a CH that is used in WCA, because it allows to extend the lifetime of nodes by relinquish the role as a CH in case of insufficient battery power. The algorithm also limits the number of nodes inside a cluster. By restricting the number of nodes catered by a cluster head helps in proper MAC functioning. The algorithm is based on the clusters' capacity and it uses the link lifetime instead of the node mobility for the maintenance procedure. The reason behind this is due to the fact that the node mobility metric does not affect the election of a CH as much as the link stability metric does. EMAC implements different mechanisms for arrival of new nodes, cluster head nodes, member nodes, merging of clusters, reelection of cluster heads.

J. *Entropy Based Clustering algorithm*

In [13], the authors have used entropy as a measure of local and mutual information available to every node. Three node parameters: mobility, energy, and degree have been used for the selection of cluster head. The method first calculates the entropy for the three node parameters. These three entropies are then combined through a simple linear model to find the total entropy of a node. The node with the lowest entropy is selected as a cluster head. The nodes gather information about their mutual mobility, energy, and degree through the exchange of beacon messages.

K. *A Fuzzy-Based Hierarchical Energy Efficient Routing Protocol (FEER)*

In [14] a fuzzy based hierarchical energy efficient routing scheme has been proposed for large scale adhoc networks that aim to prolong network's lifetime. A fuzzy logic controller has been developed that combines three node parameters residual energy, traffic and mobility to determine the node weight which denotes its suitability for acting as a cluster head.

The FEER protocol has been divided into four parts: (i) Election of cluster heads, (ii) Association of nodes with elected cluster heads, (iii) A network recovery approach to ensure fault tolerant backbone, (iv) Energy efficient routing between nodes.

FEER has been compared with the dominating set energy efficient approach [23] and it showed improved performance by prolonging the network lifetime.

In the paper, the authors have effectively depicted the use of fuzzy logic for node weight calculation. Similarly many fuzzy based cluster head selection [18] have been proposed for wireless sensor networks.

In [17] nodes are considered data objects characterized by certain attributes and the membership of the nodes to the clusters is treated in a fuzzy way.

Similarly in [22] fuzzy logic has been used to predict the link lifetime of a link connecting two nodes by taking distance and relative speed as input. The authors have introduced a new factor called δ-degree which is the number of links with lifetime greater than a particular predefined threshold.

III. DISCUSSION
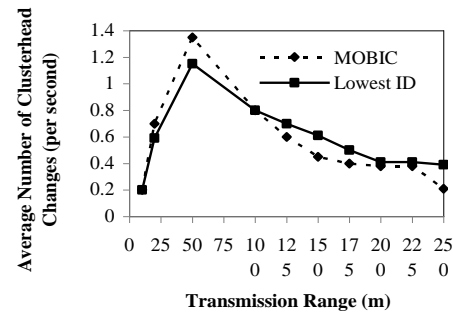
Figure 1. CH Changes: Lowest ID vs. MOBIC

Basu et.el [10] observed as in fig 1 that for moderate to high transmission ranges > 100 m, MOBIC outperforms the Lowest-ID clustering algorithm. At transmission ranges more than 125 m, the reduction in the rate of clusterhead changes is by about 0.1/*sec.* and for transmission range = 250 *m,* MOBIC yields a gain of close to 33% over Lowest-ID clustering. This can be attributed to the fact that in MOBIC the node with lowest relative mobility among all nodes get selected as clusterhead and hence have higher probability of being stable. The probability that they will change clusterheads frequently is lower as compared to the clusterheads in the Lowest-ID case.

M.Chatterjee et.el [8] made the observation as in fig.2 that the average number of clusterheads decreases with the increase in the transmission range.
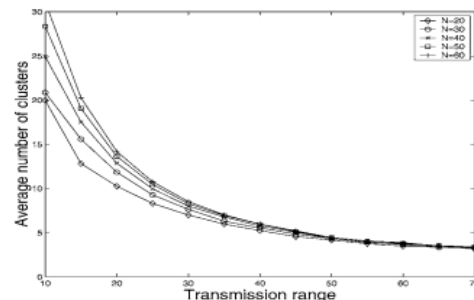
Figure 2. Average number of clusters, *max_disp* = 5
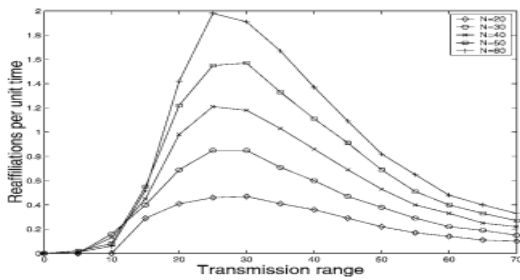
Figure 3.   Reaffiliations per unit time, *max_disp* = 5

Fig. 3 [8] shows the reaffiliations per unit time. For low transmission range, the clusters formed are either 1 or 2 node clusters which is evident from figure 2.Hence there is minimum scope of reaffiliations (fig. 3). The number of reaffiliations increases as the transmission range increases, hence increasing the cluster size   and reaches a peak when transmission range is between 25   and 30.Further increase in transmission range decrease in the reaffiliations since the nodes, in spite of their random motion, tend to stay inside the large area covered by the clusterhead.

Fig.4 [8] shows that number of dominant set updates is higher for smaller transmission range, as the transmission range increases, the number of dominant set updates decreases. This is due to the fact that at low transmission ranges the degree difference parameter of clusterhead   has high value. The weight given to this factor is 70% resulting in a higher value of total node weight for each node. This results in frequent changes in dominant set. As the transmission range increases the value of degree difference parameter of clusterhead also decreases resulting in less number of dominant set updates.
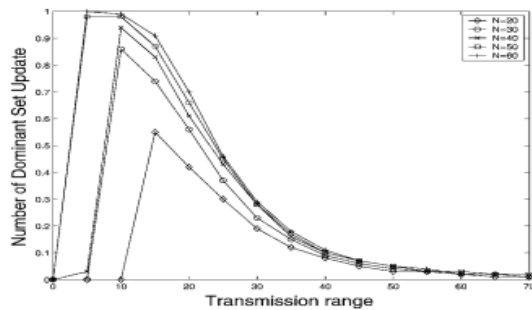
Figure 4.   Dominant set updates, *max_disp* = 5
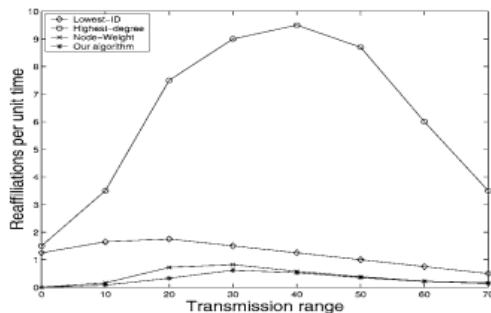
Figure 5.   Comparison of reaffiliations, *N* = 30.

Fig.5 [8] shows the relative performance of the Highest-Degree, Lowest-ID, Node-Weight heuristics and WCA in terms of the number of reaffiliations per unit time vs. transmission range where *N* = 30. The number of reaffiliations for WCA is at most half the number obtained from the Lowest-ID. The main reason is that the frequency of invoking the clustering algorithm is lower in WCA, as clustering is based on a weighing factor that results in determining stable clusters hence lesser reaffiliations. This also results in a longer duration of stability of the topology. WCA performs marginally better than the Node-Weight heuristics. WCA provides flexibility of adjusting the weighing factors according to the system needs by suitably adjusting weighing factors according to system needs.

A study of stability of the ad hoc network in terms of number of formed clusters and number of transition on each CH for different transmission ranges and network densities has been discussed for EMAC.

It can be seen from Fig. 6 [20] that for small transmission ranges, and low node density the number   of   clusters   is relatively high as most of nodes form single or two node clusters. Number of clusters decreases with increase in transmission range as more nodes come in each others ranges and form larger clusters. Similarly,   when the transmission range begins to be larger, mobile nodes tend to remain in the range of their neighbors and   the number of transitions decreases. In figure 6, when the transmission range is very small, most of nodes form one node cluster which only consists of itself. EMAC algorithm is designed so as to attempt merging of small clusters whenever possible. This causes clusterheads to switch their status to non-clustered state in order to merge with their neighbors (if any) causing the high rate of transitions in disconnected networks.
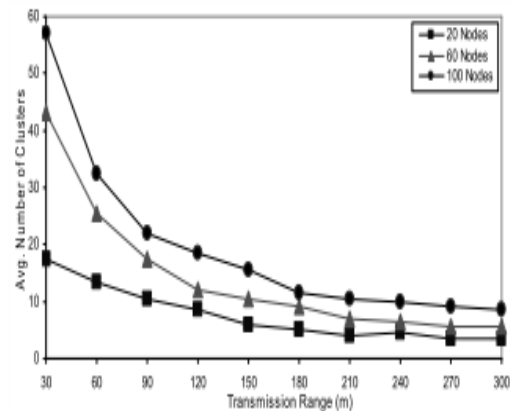
Figure 6.   Transmission range vs. Avg. Number of Clusters
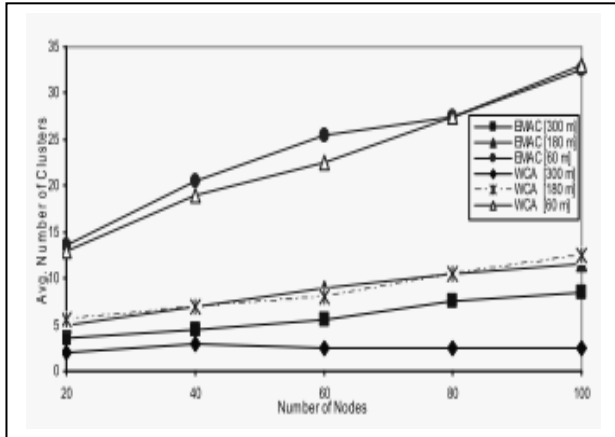
Figure 7.   Number of Nodes vs. Avg. Number of Clusters

Figure 9.   Effect of Mobility on Mean Cluster Size

Figure 8.   Number of Nodes vs. Avg. Transition Number on each CH

In figure 7 [20], it can be seen that that the performance difference is small between WCA and EMAC with respect to the average number of clusters. This is because both algorithms are variations of a local weight based clustering technique that forms one-hop clusters. As shown in figure 8, EMAC gives better performance in terms of stability as compared to WCA when the node density in the network is high. The CH of WCA algorithm relinquishes its position when another node having lower weight joins the cluster. In EMAC, the CH has to verify the suitability of a new election even if a new node having lower weight has joined the cluster.

The (α, t) clustering algorithm results Fig.9 [6] show that the mean cluster size decreases with increase in node mobility as desirable by the (α, t) criterion.

Figure 10. Effect of Mobility on Mean Cluster Size Effect of mobility on the Probability that a node is clustered with varying values of t

It is observed from fig.10,11 [6] that the nodes remain clustered even at high rates of mobility. This is because according to (α, t) criterion only those nodes form a part of cluster that have the probability of remaining connected for the specified duration.
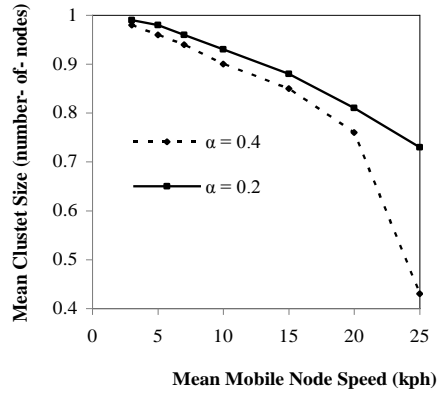
Figure 11. Effect of mobility on the Probability that a node is clustered with varying values of t
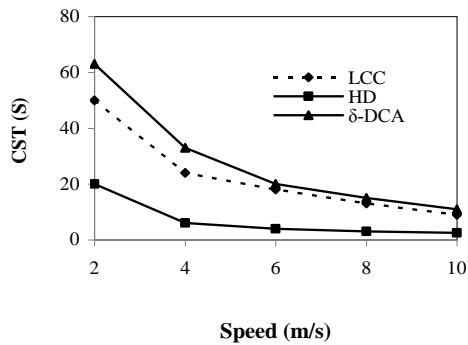
Figure 13. Average number of clusterhead changes versus transmission range.

IV. CONCLUSION

It is observed that for all clustering algorithms the number of clusters decrease with increase in transmission range, as more nodes are within range of other nodes for longer periods of time. Therefore, less number of clusters, which are larger in size, are formed, and mobility causes lesser number of nodes which are at the border to move in and out of range of each other. This results in decrease in the number of clusterhead changes.

It can be concluded from the different comparison graphs that the algorithms that are considering the different attributes in the network such as node mobility, degree of clusterhead, distance between nodes, node battery power etc. result in selecting more stable clusterheads with lesser reaffiliations and increased network lifetime. For networks with highly mobile nodes, mobility should be the critical parameter and for network with high traffic energy could be a critical parameter for clusterhead selection. Highly mobile nodes lead to more volatile clusters and should not be used as critical nodes. It can be concluded that the importance to the different parameters should be according to the network environment. Soft computing techniques can be applied to achieve clustering using existing algorithms or new algorithms and these techniques can lead to improved results.

Figure 12. Effect of mobility on Clusterhead Survival Time

Fig.12 shows Clusterhead Survival Time (CST) of δ-DCA algorithm [22] is higher than LCC and the cluster of HD algorithm.

The entropy based scheme [13] also results in forming more stable clustering topology as compared to HD and LID as can be seen from fig.13

REFERENCES

[1] M. Gerla and J.T.C. Tsai, "Multicluster, mobile, multimedia radio network,Wireless Networks" 1(3) (1995) 255–265.

[2] C.-H.R. Lin and M. Gerla, "A distributed control scheme in multi-hop packet radio networks for voice/data traffic support", in: Proceedings of IEEE GLOBECOM (1995) pp. 1238–1242.

[3] A. Ephremides, J.E. Wieselthier, D.J. Baker. "A design concept for reliable mobile radio networks with frequency hoping signaling". Proc. IEEE 75. 1987. pp. 56-73.

[4] A. Parekh. "Selecting routers in ad hoc wireless networks". Proceedings of the SBT/IEEE International Telecommunications Symposium. 1994.

[5] G. Chen, F. Nocetti, J. Gonzalez, and I. Stojmenovic, "Connectivity based k-hop clustering in wireless networks". Proceedings of the 35th Annual Hawaii International Conference on System Sciences. Vol. 7. 2002. pp. 188.3.

[6] B. McDonald and T. E Znati, "A mobility-based framework for adaptive clustering in wireless ad hoc networks", IEEE JSAC, Vol. 17, No. 8, August 1999.

[7] S. Basagni. "Distributed clustering for ad hoc networks". Proc. ISPAN'99 Int. Symp. On Parallel Architectures, Algorithms, and Networks. 1999. pp. 310-315.

[8] M. Chatterjee, S. K. Das, D. Turgut. "WCA: A weighted clustering algorithm for mobile ad hoc networks". Cluster Computing. Vol. 5. 2002. pp. 193-204.

[9] I. Er and W. Seah. "Mobility-based d-hop clustering algorithm for mobile ad hoc networks". IEEE Wireless Communications and Networking Conference. Vol. 4. 2004. pp. 2359-2364.

[10] P. Basu, N. Khan, T. D. C. Little. "A mobility based metric for clustering in mobile ad hoc networks". Proceedings of the 21st International Conference on Distributed Computing Systems. 2001. pp. 413.

[11] D. Turgut, S.K. Das, R. Elmasri, B. Turgut. "Optimizing clustering algorithm in mobile ad hoc networks using genetic algorithmic approach". Global Telecommunications Conference, IEEE. Vol. 1. 2002. pp. 62-66.

[12] D. Turgut, B. Turgut, R. Elmasri, T. V. Le. "Optimizing clustering algorithm in mobile ad hoc networks using simulated annealing". Wireless Communications and Networking, IEEE. Vol. 3. 2003. pp.1492-1497.

[13] K. Robinson, D. Turgut, M. Chatterjee. "An entropy-based clustering in mobile ad hoc networks". Proceedings of the 2006 IEEE International Conference on Networking, Sensing and Control (ICNSC). 2006. pp. 1-5.

[14] El-Hajj, W.; Kountanis, D.; Al-Fuqaha, A.; Guizani, M. "A fuzzy-based hierarchical energy efficient routing protocol for large scale mobile ad hoc networks (FEER)." IEEE International Conference on Volume 8, Issue , June 2006 Page(s):3585 – 3590.

[15] C.-C. Chiang, H.-K. Wu, W. Liu, and M. Gerla. "Routing in clustered multihop, mobile wireless networks with fading channel." , IEEE Singapore International Conference on Networks (SICON), pages 197-211, Apr. 1997.

[16] M. Gerla and J. T.-C. Tsai, "Multicluster, mobile, multimedia radio network," ACM-Baltzer J. Wireless Networks journal 95,, vol. 1, no. 3,oct 1995, pp. 255–265.

[17] Jörg Habetha and Bernhard Walke, "Fuzzy rule-based mobility and load management for self-organizing wireless networks" International Journal of Wireless Information Networks, Vol. 9, No. 2, April 2002 (© 2002)

[18] Indranil Gupta Denis Riordan Srinivas Sampalli, "Cluster-head eection using fuzzy logic for wireless sensor networks", Proceedings of the 3rd Annual Communication Networks and Services Research Conference (CNSR'05) IEEE 2005.

[19] Nagaraju Uppu, B.V.S.S. Subhramanyam and Ramamurthy Garimella, "An Energy efficient technique to polong network lifetime of adhoc sensor networks(ETPNL)" , IETE Technical Review Vol.25, Issue 4, Jul-Aug 2008, pp. 154-160.

[20] Zouhair El-Bazzal, Michel Kadoch, Basile L. Agba, François Gagnon and Maria Bennani, "An efficient management algorithm for clustering in mobile ad hoc network", in Proceedings of the ACM international workshop on Performance monitoring, measurement, and evaluation of heterogeneous wireless and wired networks, 2006, pp. 5 - 31 .

[21] A. B. McDonald, T. F. Znati. "Design and simulation of a distributed dynamic clustering algorithm for multimode routing in wireless ad hoc networks". SIMULATION. Vol. 78. 2002. pp. 408-422.

[22] ZHAO Chun-xiao, WANG Guang-xing, "Fuzzy control-based clustering strategy in MANET", Proceedings of the 5th World Congress 2004.

[23] J. Wu et al., "On calculating power-aware connected dominating sets for efficient routing in ad hoc wireless networks," J. Commun. And Networks, vol. 4, no. 1, Mar. 2002, pp. 59–70.

[24] M. Chatterjee, S.K. Das and D. Turgut, "An on-demand weighted clustering algorithm (WCA) for ad hoc networks", in: Proceedings of IEEE GLOBECOM 2000, San Francisco, November 2000, pp. 1697–1701.

[25] S. Basagni, I. Chlamtac and A. Farago, "A generalized clustering algorithm for peer-to-peer networks", in: Proceedings of Workshop on Algorithmic Aspects of Communication (satellite workshop of ICALP), July 1997.

[26] C.-H.R. Lin and M. Gerla, "A distributed architecture for multimedia in dynamic wireless networks", in: Proceedings of IEEE GLOBECOM (1995) pp. 1468–1472.

[27] S. Kirkpatrick , C.D. Gelatt Jr. , and M.P. Vecchi, "Optimization by Simulated Annealing" , Science 220, 1983, pp. 671 – 680.

Survey of Nearest Neighbor Techniques

Nitin Bhatia (Corres. Author)
Department of Computer Science
DAV College
Jalandhar, INDIA
n_bhatia78@yahoo.com

Vandana
SSCS
Deputy Commissioner's Office
Jalandhar, INDIA
vandana_ashev@yahoo.co.in

Abstract— **The nearest neighbor (NN) technique is very simple, highly efficient and effective in the field of pattern recognition, text categorization, object recognition etc. Its simplicity is its main advantage, but the disadvantages can't be ignored even. The memory requirement and computation complexity also matter. Many techniques are developed to overcome these limitations. NN techniques are broadly classified into structure less and structure based techniques. In this paper, we present the survey of such techniques. Weighted kNN, Model based kNN, Condensed NN, Reduced NN, Generalized NN are structure less techniques whereas k-d tree, ball tree, Principal Axis Tree, Nearest Feature Line, Tunable NN, Orthogonal Search Tree are structure based algorithms developed on the basis of kNN. The structure less method overcome memory limitation and structure based techniques reduce the computational complexity.**

Keywords- Nearest neighbor (NN), kNN, Model based kNN, Weighted kNN, Condensed NN, Reduced NN.

I. INTRODUCTION

The nearest neighbor (NN) rule identifies the category of unknown data point on the basis of its nearest neighbor whose class is already known. This rule is widely used in pattern recognition [13, 14], text categorization [15-17], ranking models [18], object recognition [20] and event recognition [19] applications.

T. M. Cover and P. E. Hart purpose k-nearest neighbor (kNN) in which nearest neighbor is calculated on the basis of value of k, that specifies how many nearest neighbors are to be considered to define class of a sample data point [1]. T. Bailey and A. K. Jain improve kNN which is based on weights [2]. The training points are assigned weights according to their distances from sample data point. But still, the computational complexity and memory requirements remain the main concern always. To overcome memory limitation, size of data set is reduced. For this, the repeated patterns, which do not add extra information, are eliminated from training samples [3-5]. To further improve, the data points which do not affect the result are also eliminated from training data set [6]. Besides the time and memory limitation, another point which should be taken care of, is the value of k, on the basis of which category of the unknown sample is determined. Gongde Guo selects the value of k using model based approach [7]. The model proposed automatically selects the value of k. Similarly, many improvements are proposed to improve speed of classical kNN using concept of ranking [8], false neighbor information [9], clustering [10]. The NN training data set can be structured using various techniques to improve over memory limitation of

kNN. The kNN implementation can be done using ball tree [21, 22], k-d tree [23], nearest feature line (NFL) [24], tunable metric [26], principal axis search tree [28] and orthogonal search tree [29]. The tree structured training data is divided into nodes, whereas techniques like NFL and tunable metric divide the training data set according to planes. These algorithms increase the speed of basic kNN algorithm.

II. NEAREST NEIGHBOR TECHNIQUES

Nearest neighbor techniques are divided into two categories: 1) Structure less and 2) Structure Based.

A. Structure less NN techniques

The k-nearest neighbor lies in first category in which whole data is classified into training data and sample data point. Distance is evaluated from all training points to sample point and the point with lowest distance is called nearest neighbor.

This technique is very easy to implement but value of k affects the result in some cases. Bailey uses weights with classical kNN and gives algorithm named weighted kNN (WkNN) [2]. WkNN evaluates the distances as per value of k and weights are assigned to each calculated value, and then nearest neighbor is decided and class is assigned to sample data point. The Condensed Nearest Neighbor (CNN) algorithm stores the patterns one by one and eliminates the duplicate ones. Hence, CNN removes the data points which do not add more information and show similarity with other training data set. The Reduced Nearest Neighbor (RNN) is improvement over CNN; it includes one more step that is elimination of the patterns which are not affecting the training data set result. The another technique called Model Based kNN selects similarity measures and create a 'similarity matrix' from given training set. Then, in the same category, largest local neighbor is found that covers large number of neighbors and a data tuple is located with largest global neighborhood. These steps are repeated until all data tuples are grouped. Once data is formed using model, kNN is executed to specify category of unknown sample. Subash C. Bagui and Sikha Bagui [8] improve the kNN by introducing the concept of ranks. The method pools all the observations belonging to different categories and assigns rank to each category data in ascending order. Then observations are counted and on the basis of ranks class is assigned to unknown sample. It is very much useful in case of multi-variants data. In Modified kNN, which is modification of WkNN validity of all data samples in the training data set is computed, accordingly weights are assigned and then validity and weight both together set basis for classifying the class of

the sample data point. Yong zeng, Yupu Zeng and Liang Zhou define the new concept to classify sample data point. The method introduces the pseudo neighbor, which is not the actual nearest neighbor; but a new nearest neighbor is selected on the basis of value of weighted sum of distances of kNN of unclassified patterns in each class. Then Euclidean distance is evaluated and pseudo neighbor with greater weight is found and classified for unknown sample. In the technique purposed by Zhou Yong [11], Clustering is used to calculate nearest neighbor. The steps include, first of all removing the samples which are lying near to the border, from training set. Cluster each training set by k value clustering and all cluster centers form new training set. Assign weight to each cluster according to number of training samples each cluster have.

B. Structure based NN techniques

The second category of nearest neighbor techniques is based on structures of data like Ball Tree, k-d Tree, principal axis Tree (PAT), orthogonal structure Tree (OST), Nearest feature line (NFL), Center Line (CL) etc. Ting Liu introduces the concept of Ball Tree. A ball tree is a binary tree and constructed using top down approach. This technique is improvement over kNN in terms of speed. The leaves of the tree contain relevant information and internal nodes are used to guide efficient search through leaves. The k-dimensional trees divide the training data into two parts, right node and left node. Left or right side of tree is searched according to query records. After reaching the terminal node, records in terminal node are examined to find the closest data node to query record. The concept of NFL given by Stan Z.Li and Chan K.L. [24] divide the training data into plane. A feature line (FL) is used to find

nearest neighbor. For this, FL distance between query point and each pair of feature line is calculated for each class. The resultant is set of distances. The evaluated distances are sorted into ascending order and the NFL distance is assigned as rank 1. An improvement made over NFL is Local Nearest Neighbor which evaluates the feature line and feature point in each class, for points only, whose corresponding prototypes are neighbors of query point. Yongli Zhou and Changshui Zhang introduce [26] new metric for evaluating distances for NFL rather than feature line. This new metric is termed as "Tunable Metric". It follows the same procedure as NFL but at first stage it uses tunable metric to calculate distance and then implement steps of NFL. Center Based Nearest Neighbor is improvement over NFL and Tunable Nearest Neighbor. It uses center base line (CL) that connects sample point with known labeled points. First of all CL is calculated, which is straight line passing through training sample and center of class. Then distance is evaluated from query point to CL, and nearest neighbor is evaluated. PAT permits to divide the training data into efficient manner in term of speed for nearest neighbor evaluation. It consists of two phases 1) PAT Construction 2) PAT Search. PAT uses principal component analysis (PCA) and divides the data set into regions containing the same number of points. Once tree is formed kNN is used to search nearest neighbor in PAT. The regions can be determined for given point using binary search. The OST uses orthogonal vector. It is an improvement over PAT for speedup the process. It uses concept of "length (norm)", which is evaluated at first stage. Then orthogonal search tree is formed by creating a root node and assigning all data points to this node. Then left and right nodes are formed using pop operation.

TABLE I.   COMPARISON OF NEAREST NEIGHBOR TECHNIQUES

Sr No	Technique	Key Idea	Advantages	Disadvantages	Target Data
1.	k Nearest Neighbor (kNN) [1]	Uses nearest neighbor rule	1. training is very fast 2. Simple and easy to learn 3. Robust to noisy training data 4.Effective if training data is large	1. Biased by value of k 2.Computation Complexity 3.Memory limitation 4. Being a supervised learning lazy algorithm i.e. runs slowly 5. Easily fooled by irrelevant attributes	large data samples
2.	Weighted k nearest neighbor (WkNN) [2]	Assign weights to neighbors as per distance calculated	1. Overcomes limitations of kNN of assigning equal weight to k neighbors implicitly. 2. Use all training samples not just k. 3. Makes the algorithm global one	1. Computation complexity increases in calculating weights 2. Algorithm runs slow	Large sample data
3.	Condensed nearest neighbor (CNN) [3,4,5]	Eliminate data sets which show similarity and do not add extra information	1. Reduce size of training data 2. Improve query time and memory requirements 3.Reduce the recognition rate	1. CNN is order dependent; it is unlikely to pick up points on boundary. 2. Computation Complexity	Data set where memory requirement is main concern
4.	Reduced Nearest Neigh (RNN) [6]	Remove patterns which do not affect the training data set results	1. Reduce size of training data and eliminate templates 2. Improve query time and memory requirements 3.Reduce the recognition rate	1.Computational Complexity 2.Cost is high 3.Time Consuming	Large data set
5.	Model based k nearest neighbor (MkNN) [7]	Model is constructed from data and classify new data using model	1. More classification accuracy 2.Value of k is selected automatically 3.High efficiency as reduce number of data points	1.Do not consider marginal data outside the region	Dynamic web mining for large repository
6.	Rank nearest neighbor (kRNN) [8]	Assign ranks to training data for each category	1.Performs better when there are too much variations between features 2.Robust as based on rank	1.Multivariate kRNN depends on distribution of the data	Class distribution of Gaussian nature

			3.Less computation complexity as compare to kNN		
7.	Modified k nearest neighbor (MkNN) [10]	Uses weights and validity of data point to classify nearest neighbor	1.Partially overcome low accuracy of WkNN 2.Stable and robust	1.Computation Complexity	Methods facing outlets
8.	Pseudo/Generalized Nearest Neighbor (GNN) [9]	Utilizes information of n-1 neighbors also instead of only nearest neighbor	1.uses n-1 classes which consider the whole training data set	1.does not hold good for small data 2.Computationa;l complexity	Large data set
9.	Clustered k nearest neighbor [11]	Clusters are formed to select nearest neighbor	1.Overcome defect of uneven distributions of training samples 2.Robust in nature	1.Selection of threshold parameter is difficult before running algorithm 2.Biased by value of k for clustering	Text Classification
10.	Ball Tree k nearest neighbor (KNS1) [21,22]	Uses ball tree structure to improve kNN speed	1.Tune well to structure of represented data 2.Deal well with high dimensional entities 3.Easy to implement	1.Costly insertion algorithms 2.As distance increases KNS1 degrades	Geometric Learning tasks like robotic, vision, speech, graphics
11.	k-d tree nearest neighbor (kdNN) [23]	divide the training data exactly into half plane	1.Produce perfectly balanced tree 2.Fast and simple	1.More computation 2.Require intensive search 3.Blindly slice points into half which may miss data structure	organization of multi dimensional points
12.	Nearest feature Line Neighbor (NFL) [24]	take advantage of multiple templates per class	1.Improve classification accuracy 2.Highly effective for small size 3.utilises information ignored in nearest neighbor i.e. templates per class	1.Fail when prototype in NFL is far away from query point 2.Computations Complexity 3.To describe features points by straight line is hard task	Face Recognition problems
13.	Local Nearest Neighbor [25]	Focus on nearest neighbor prototype of query point	1.Cover limitations of NFL	1.Number of Computations	Face Recognition
14.	Tunable Nearest Neighbor (TNN) [26]	A tunable metric is used	1.Effective for small data sets	1.Large number of computations	Discrimination problems
15.	Center based Nearest Neighbor (CNN) [27]	A Center Line is calculated	1.Highly efficient for small data sets	1. Large number of computations	Pattern Recognition
16.	Principal Axis Tree Nearest Neighbor (PAT) [28]	Uses PAT	1.Good performance 2.Fast Search	1.Computation Time	Pattern Recognition
17.	Orthogonal Search Tree Nearest Neighbor [29]	Uses Orthogonal Trees	1.Less Computation time 2.Effective for large data sets	1.Query time is more	Pattern Recognition

III. CONCLUSION

We compared the nearest neighbor techniques. Some of them are structure less and some are structured base. Both kinds of techniques are improvements over basic kNN techniques. Improvements are proposed by researchers to gain speed efficiency as well as space efficiency. Every technique hold good in particular field under particular circumstances.

REFERENCES

[1] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE Trans. Inform. Theory, Vol. IT-13, pp 21-27, Jan 1967.

[2] T. Bailey and A. K. Jain, "A note on Distance weighted k-nearest neighbor rules", IEEE Trans. Systems, Man Cybernatics, Vol.8, pp 311-313, 1978.

[3] K. Chidananda and G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighbor", IEEE Trans. Information Theory, Vol IT- 25 pp. 488-490, 1979.

[4] F Angiulli, "Fast Condensed Nearest Neighbor", ACM International Conference Proceedings, Vol 119, pp 25-32.

[5] E Alpaydin, "Voting Over Multiple Condensed Nearest Neighbors", Artificial Intelligent Review 11:115-132, 1997.

[6] Geoffrey W. Gates, "Reduced Nearest Neighbor Rule", IEEE Trans Information Theory, Vol. 18 No. 3, pp 431-433.

[7] G. Guo, H. Wang, D. Bell, "KNN Model based Approach in Classification", Springer Berlin Vol 2888.

[8] S. C. Bagui, S. Bagui, K. Pal, "Breast Cancer Detection using Nearest Neighbor Classification Rules", Pattern Recognition 36, pp 25-34, 2003.

[9] Y. Zeng, Y. Yang, L. Zhou, "Pseudo Nearest Neighbor Rule for Pattern Recognition", Expert Systems with Applications (36) pp 3587-3595, 2009.

[10] H. Parvin, H. Alizadeh and B. Minaei, "Modified k Nearest Neighbor", Proceedings of the world congress on Engg. and computer science 2008.

[11] Z. Yong, "An Improved kNN Text Classification Algorithm based on Clustering", Journal of Computers, Vol. 4, No. 3, March 2009.

[12] Djouadi and E. Bouktache, "A Fast Algorithm for Nearest Neighbor Classification", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19. No. 3, 1997.

[13] V.Vaidehi, S. Vasuhi, "Person Authentication using Face Recognition", Proceedings of the world congress on engg and computer science, 2008.

[14] Shizen, Y. Wu, "An Algorithm for Remote Sensing Image Classification based on Artificial Immune b-cell Network", Springer Berlin, Vol 40.

[15] G. Toker, O. Kirmemis, "Text Categorization using k Nearest Neighbor Classification", Survey Paper, Middle East Technical University.

[16] Y. Liao, V. R. Vemuri, "Using Text Categorization Technique for Intrusion detection", Survey Paper, University of California.

[17] E. M. Elnahrawy, "Log Based Chat Room Monitoring Using Text Categorization: A Comparative Study", University of Maryland.

[18] X. Geng et. al, "Query Dependent Ranking Using k Nearest Neighbor", SIGIR, 2008.

[19] Y. Yang and T. Ault, "Improving Text Categorization Methods for event tracking", Carnegie Mellon University.

[20] F. Bajramovie et. al "A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition", ACIVS 2006, LNCS 4179, pp 1186-1197.

[21] T. Liu, A. W. Moore, A. Gray, "New Algorithms for Efficient High Dimensional Non-Parametric Classification", Journal of Machine Learning Research, 2006, pp 1135-1158.

[22] S. N. Omohundro, "Five Ball Tree Construction Algorithms", 1989, Technical Report.

[23] R. F Sproull, "Refinements to Nearest Neighbor Searching", Technical Report, International Computer Science, ACM (18) 9, pp 507-517.

[24] S. Z Li, K. L. Chan, "Performance Evaluation of The NFL Method in Image Classification and Retrieval", IEEE Trans On Pattern Analysis and Machine Intelligence, Vol 22, 2000.

[25] W. Zheng, L. Zhao, C. Zou, "Locally Nearest Neighbor Classifier for Pattern Classification", Pattern Recognition, 2004, pp 1307-1309.

[26] Y. Zhou, C. Zhang, "Tunable Nearest Neighbor Classifier", DAGM 2004, LNCS 3175, pp 204-211.

[27] Q. B. Gao, Z. Z. Wang, "Center Based Nearest Neighbor Class", Pattern Recognition, 2007, pp 346-349.

[28] Y. C. Liaw, M. L. Leou, "Fast Exact k Nearest Neighbor Search using Orthogonal Search Tree", Pattern Recognition 43 No. 6, pp 2351-2358.

[29] J. Mcname, "Fast Nearest Neighbor Algorithm based on Principal Axis Search Tree", IEEE Trans on Pattern Analysis and Machine Intelligence, Vol 23, pp 964-976.

Time Domain Analysis based Fault Diagnosis Methodology for Analog Circuits-A Comparative Study of Fuzzy and Neural Classifier Performance

V. Prasannamoorthy[1], R. Bharat Ram[2], V. Manikandan[3], N. Devarajan[4]

[1,2,4] *Department of Electrical Engineering, Government College of Technology*
Coimbatore, India

[3]*Department of Electrical Engineering, Coimbatore Institute of Technology*
Coimbatore, India

[1] `prasanna_gct1995@yahoomail.com`

Abstract— **In this paper, we attempt to diagnose the occurrence of faults in analog electronic circuits based upon variations in time domain specifications corresponding to the circuit condition under consideration relative to the fault free circuit. To achieve this, both a fuzzy as well as a neural classifier have been utilized to operate with the fault dictionary data as base. Through this process, a general comparison is drawn out between the performance of either route in dealing with fault diagnosis of circuits. An illustrative example is considered, on which both the fuzzy and neural algorithms are tested, and their performance in fault diagnosis is compared. Further, the suitability of the fuzzy and neural techniques to various kinds of diagnosis problems depending upon the nature of data available is also discussed.**

Keywords—Fault diagnosis, fuzzy logic system, neural networks,Sallen-key Bandpass filter.

I. INTRODUCTION

The identification of faults in any analog circuit is very useful and, in a few instances, an inevitable measure in ensuring competent performance of the circuit. In general, the analog diagnosis approaches can be categorized into two [1], namely-simulation-after-test (SAT) and simulation-before-test (SBT). The simulation-after-test [2]-[4] approach involves the computation of various circuit parameters from the operational circuit and fault identification is carried out using these parameters, assuming that each measurement is independent of the other. This method is avoided due to the increase in process time with increase in the size of the circuit, in addition to issues concerning non-linear circuits. On the other hand, a useful alternative is found in the simulation-before-test approach which appreciably reduces the time taken for fault diagnosis. In the SBT approach [5]-[9], a predefined set of test stimuli are used to extract certain signatures from the Circuit-Under-Test (CUT) that are unique to each faulty condition. These signatures can then be suitably systematized to create a "fault dictionary", which is then checked for redundancies that may result in masking of certain faults. Evidently, the parameters chosen to pose as signatures must be quantities that are observable for all conditions of the circuit.

Both the above-mentioned approaches are fairly procedural in nature and do not necessitate the prerequisite of an intuitional knowledge of the functioning of the CUT. Constant supervision of the circuit is entailed to ensure stable performance over an extended period of time. The identification of faults in systems is often a combination of fault detection and isolation, necessarily in the same order, which is commonly known as FDI [10]-[11]. Early detection of faults in a circuit can greatly assist in maintenance of the system by avoiding possibly harmful damage borne out of the fault. Occasionally, a circuit may so damaged that it might assume an unstable state, making it impossible to extract signatures from it that might help in identifying the fault. In other cases, a fault might just be too critical or dangerous to be provoked for the sake of obtaining a signature.

Analog fault diagnosis is inherently complicated by poor mathematical models, component tolerances, nonlinear behaviour of components, and limited accessibility to internal nodes of the circuit under test. In this paper, we state the results of a comparative study of the performance of fuzzy and neural routes in the detection and identification of faults in an analog electronic circuit using the Simulation-Before-Test approach. This was achieved by taking into consideration the variations in time-domain response parameters pertaining to the transient of the CUT for a step input. A comprehensive fault dictionary was prepared from all the possible values of the parameters corresponding to each state of the circuit, which was then effectively utilized to construct a classifier capable of identifying the various faulty configurations of the CUT.

II. GENERALIZED ALGORITHM

The fault diagnosis methodology, involving either a fuzzy or a neural system, may be divided into five distinct steps as follows:

Step I: Formulation of transfer function of the circuit under test assuming nominal values of all components in the circuit.

Step II: Simulation of time-domain response of the circuit when a unit step signal is given as input, for possible combinations of component faults.

Step III: Determination of time-domain response parameters, namely, settling time and peak amplitude for each time response plot.

Step IV: Suitable pre-processing of available data and design of classifier.

Step V: Isolation of faults by using the signatures extracted from a suspicious circuit by feeding its time response parameters to the classifier obtained in step IV.

III. TEST CIRCUIT

To test the performance of the fuzzy and neural techniques, we chose the Sallen-Key bandpass filter circuit [12] shown in Fig. 1. The Sallen-Key bandpass filter is a second order active filter, which is greatly appreciated for its simplicity of design. The filter section shown in Fig. 1 can also be cascaded to form second order filter stages, resulting in larger order filters. The op-amp provides buffering between filter stages, so that each stage can be designed independently. This circuit is suitable for filters which have complex conjugate poles. When implementing a particular transfer function, a designer will typically find all the poles, and group them into real poles and complex conjugate pairs. Each of the complex conjugate pole pairs are then implemented with a Sallen-Key filter, and the circuits are cascaded together to form the complete filter.
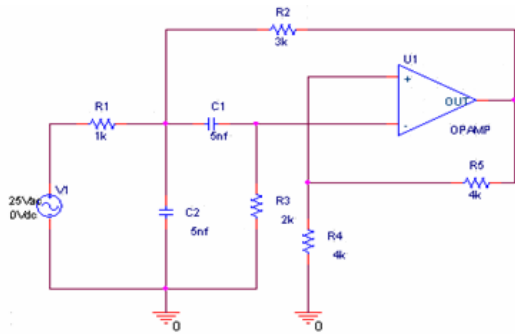
Fig. 1 Sallen-Key Bandpass Filter circuit

The transfer function of the Sallen-Key bandpass filter circuit is

$$G(s) = \frac{sA_0 G_1 G_1}{s^2 C_1 C_2 + s(G_3 C_1 + G_2 C_2 + G_1 C_1 (1 - A_0)) + G_3 (G_1 + G_2)}$$

where

$$A_0 = 1 + \left(\frac{R_5}{R_4}\right), \qquad G_1 = \left(\frac{1}{R_1}\right),$$

$$G_2 = \left(\frac{1}{R_2}\right), \qquad G_3 = \left(\frac{1}{R_3}\right)$$

The circuit shown in Fig. 1 has the component values that correspond to the nominal frequency of operation of 25 kHz.

IV. SIGNATURE EXTRACTION

In order to obtain the fault signatures for each fault condition, the values of components are varied to +50% or -50% of their nominal values shown in Fig. 1 and the circuit is excited using a unit step signal as input, thus enabling to construct the fault dictionary. While the values of the supposedly faulty components are manipulated, it is ensured

that the remaining components maintain their nominal values as in the original circuit. Distinct variations in response may be seen for every faulty configuration, as shown in Fig. 2, 3 and 4. The two required time-domain response parameters, settling time and peak amplitude are noted for all the response curves. These, being characteristic to a particular configuration, are used as the fault signatures.
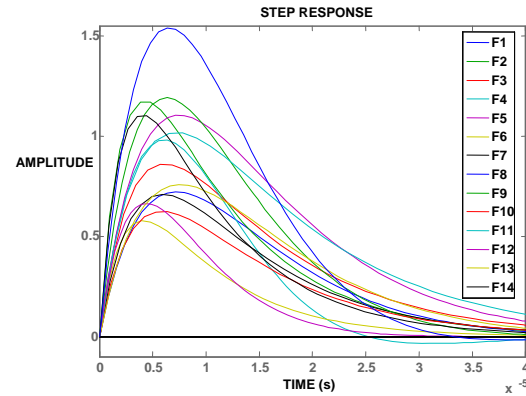
Fig. 2 Step response curves for single faults
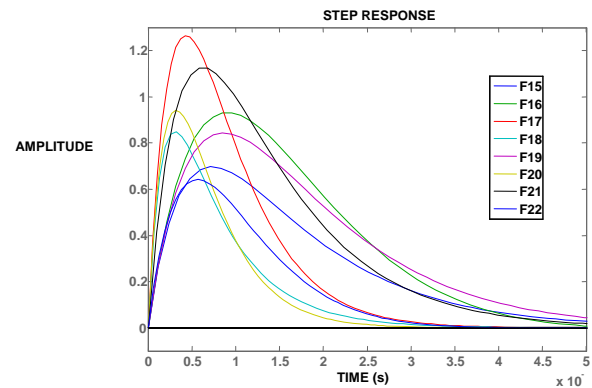
Fig. 3 Step response curves for double faults
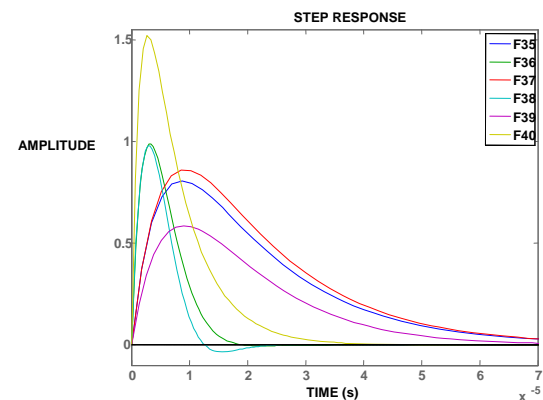
Fig. 4 Step response curves for multiple faults

There are a total 2n number of single faults for which the step response plot in shown in Fig. 2, whereas there are n(n-1)/2 number of double faults, n(n-1)/3 triple faults and n(n-1)/4 quadruple faults for which the step response curves are shown in Fig. 3 and Fig. 4. Other combinations of component faults get eliminated automatically due to repetitions of the same in the fault dictionary, the formulae stated above holding good in determining the final count of fault conditions post elimination.

accordingly, fault dictionaries are created. A general rule of thumb is that once the fault signatures have been collected and organized into a fault dictionary, the data must be optimized by eliminating signatures that bring about masking of faults whose signatures match. However, in the case of a fault dictionary made up of time-domain response parameters, the data is found to be free of redundancies hence making it feasible to identify each faulty configuration on the basis of variations in two parameters alone.

TABLE I

TIME DOMAIN SPECIFICATIONS FOR SINGLE FAULTS

Fault ID	Faulty components	Settling time *10^-5 sec	Peak Amplitude
F1	R1↑	4.3834	0.7211
F2	R1↓	4.2557	1.1709
F3	R2↑	5.2859	0.8600
F4	R2↓	3.8230	0.9788
F5	R3↑	5.1273	1.1024
F6	R3↓	3.6507	0.5770
F7	R4↑	4.7622	0.7089
F8	R4↓	3.0360	1.5381
F9	R5↑	3.6338	1.1908
F10	R5↓	4.9734	0.6228
F11	C1↑	6.1093	1.0152
F12	C1↓	2.6625	0.6645
F13	C2↑	4.7794	0.7566
F14	C2↓	3.9499	1.1012

TABLE II

TIME DOMAIN SPECIFICATIONS FOR DOUBLE FAULTS

Fault ID	Faulty components	Settling time *10^-5 sec	Peak Amplitude
F15	R1↑,R2↑	5.7921	0.6968
F16	R1↓,R2↓	3.0089	1.2654
F17	R1↑,R3↑	4.6712	0.9312
F18	R1↓,R3↓	2.9020	0.8483
F19	R1↑,C1↑	6.0179	0.8436
F20	R1↓,C1↓	2.3671	0.9403
F21	R2↑,R5↑	4.8269	1.1242
F22	R2↓,R5↓	3.2166	0.6416
F23	R3↑,R5↑	4.1115	1.5028
F24	R3↓,R5↓	3.9499	0.413
F25	R3↑,C1↑	7.4602	1.2456
F26	R3↓,C1↓	2.4867	0.4152
F27	R3↑,C2↑	5.4507	0.9526
F28	R3↓,C2↓	3.0668	0.7383
F29	R5↑,C1↑	5.1273	1.3780
F30	R5↓,C1↓	3.0089	0.4745
F31	R5↑,C2↑	4.1115	1.0019
F32	R5↓,C2↓	4.5108	0.7619
F33	R2↑,C2↑	5.8528	0.7379
F34	R2↓,C2↓	1.993	1.2744

TABLE III

TIME DOMAIN SPECIFICATIONS FOR TRIPLE FAULTS

Fault ID	Faulty components	Settling time *10^-5 sec	Peak Amplitude
F35	R1↑,R2↑,C1↑	7.9128	0.8071
F36	R1↓,R2↓,C1↓	1.6563	0.9866
F37	R3↑,R4↑,C2↑	6.1429	0.7571
F38	R3↓,R4↓,C2↓	2.4867	1.2456
F39	R1↑,R2↑,C2↑	6.6527	0.586
F40	R1↓,R2↓,C2↓	2.9279	1.5191
F41	R1↑,R3↑,C1↑	6.5752	1.0816
F42	R1↓,R3↓,C1↓	1.7798	0.6463
F43	R2↑,R3↑,C1↑	9.2037	1.1879
F44	R2↓,R3↓,C1↓	1.6083	0.4277
F45	R1↑,C1↑,C2↑	6.5752	0.7211
F46	R1↓,C1↓,C2↓	2.1278	1.1709
F47	R2↑,R5↑,C1↑	6.7579	1.2896
F48	R2↓,R5↓,C1↓	1.9938	0.4779

TABLE IV

TIME DOMAIN SPECIFICATIONS FOR QUADRUPLE FAULTS

Fault ID	Faulty components	Settling time *10^-5 sec	Peak Amplitude
F49	R1↑,R2↑,R3↑,R5↑	5.4507	1.1908
F50	R1↓,R2↓,R3↓,R5↓	2.4867	0.6228
F51	R1↑,R2↑,R3↑,C1↑	9.1640	1.0152
F52	R1↓,R2↓,R3↓,C1↓	1.3313	0.6645
F53	R1↑,R4↑,C1↑,C2↑	7.4992	0.5693
F54	R1↓,R4↓,C1↓,C2↓	1.7798	1.9389
F55	R2↑,R4↑,C1↑,C2↑	8.362	0.6978
F56	R2↓,R4↓,C1↓,C2↓	3.3142	2.0385
F57	R4↑,R5↑,C1↑,C2↑	6.487	0.888
F58	R4↓,R5↓,C1↓,C2↓	2.1623	0.888
F59	R1R2R3R4R5C1C	9.7305	0.888
F60	R1R2R3R4R5C1C	1.0812	0.888
F61	Fault Free	4.3247	0.888

From the above shown curves, the values of settling time and peak value for each fault condition are noted down and

The fault dictionaries for single component fault, double component fault, triple component fault and other multiple component faults are presented in the tables I, II, III and IV respectively. An upward arrow indicates a deviation of 50% above the nominal value, whereas a downward arrow indicates a 50% decrement. The fault free condition is also tabulated and is given as ID F61.

V. NEURAL FAULT DIAGNOSIS SYSTEM

A. Preprocessor

A complete fault dictionary containing all feasible conditions cannot be generated because of the presence of noise. This problem is solved by giving inputs to the neural network in terms of bits—a "0" is assigned if the value observed for a specific test frequency is out of bounds; a "1" is assigned if the value observed for a specific test frequency is within bounds. That is, if $X_L \leq X_m \leq X_H$ implies ANN input = 1, else ANN input = 0.

B. ANN Classifier

Artificial neural networks provide an adaptive mechanism for the task of pattern classification [13]. They are capable of reliable classification even in undesirable environments characterized by ill-defined models, noisy inputs and nonlinearity. A comparison of neural network architectures has already been undertaken by S. Hsu, et al [14]. In this paper, the back propagation network (BPN) structure has been chosen for the classification task.

A typical BPN has two or three layers of interconnecting weights. Fig. 5 shows a standard two-layer BPN network topology. Each input node is connected to a hidden layer node. Each hidden node is connected to an output node in a similar fashion. This makes the BPN a fully connected network topology.
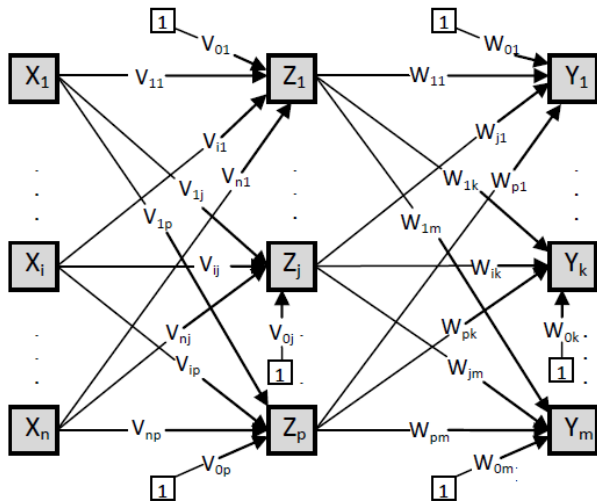
Fig. 5 Two layer BPN network topology

Here, $X_1 \ldots X_i \ldots X_n$ indicate the input neurons, $Z_1 \ldots Z_j \ldots Z_p$ the hidden layer neurons, and $Y_1 \ldots Y_k \ldots Y_m$ the output neurons of the artificial neural network. V_{ij} is the weight from i^{th} input neuron to j^{th} hidden neuron; W_{jk} is the weight from j^{th} hidden neuron to k^{th} output neuron; V_{oj} is the weight from bias to j^{th} hidden neuron, and W_{0k} is the weight from bias to k^{th} output neuron. The supervised learning in BPN takes place by propagating the node activation function of input pattern to output nodes. These outputs are compared

with the desired target values, and an error signal (δ) is produced. The network weights are adapted so as to minimize the error. The generalized delta rule does the weight adaptation given by $\Delta_p \omega_{i\ j} = \sum \delta_{pj} x_{pi}$, where $\sum$ is the learning rate, δ_{pj} is the error at the jth node due to pattern p; x_{pi} is the ith element of the output pattern p. The error signal for the output node is $\delta_{pj} = (t_{pj} - o_{pj}) f_j (net_{pj})$, where t_{pj} and o_{pj} are the target and output values respectively.

The number of nodes in a layer and the activation function will affect the learning rate, the computational complexity, and the usefulness of the network for a specific problem wherein the best results always come from intuition and experience.

The number of neurons in the input layer is 10 and the number of neurons in the hidden single layer is 21. So the ANN structure boils down to 10:21:1. The ANN is adaptively trained to update the weights and the bias by gradient descent method by the mean-square-error performance.
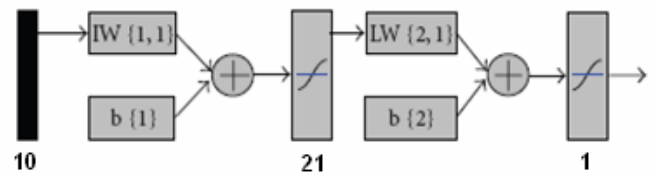
Fig. 6 Classifier for test circuit

The classifier structure for the circuit is shown in Fig. 6 In the classifier, the first block indicates the input layer comprising 10 neurons, the center block indicates the hidden layer comprising 21 neurons, and the last block indicating the output layer comprising of 1neuron, respectively. The blocks
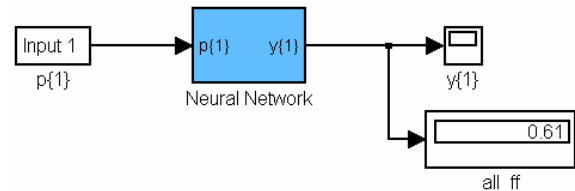
Fig. 7 Neural network Classifier for single fault
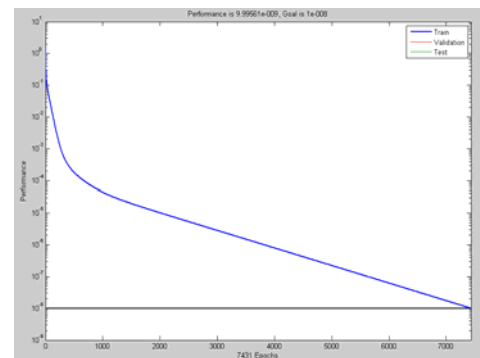
Fig. 8 Performance plot for single fault

TABLE V

ANN OUPUT FOR RANDOM FAULT CONDITIONS

Component Values (R in kΩ and C in nF)							Classifier Inputs		Fault ID
R1	R2	R3	R4	R5	C1	C2	Settling Time	Peak Amplitude	
1	3	**1**	4	4	5	5	3.6507	0.5770	F6
1	**1.5**	2	4	4	5	**2.5**	1.993	1.2744	F34
1	3	2	4	4	5	**2.5**	3.9499	1.1012	F14
0.5	**1.5**	**1**	4	4	**2.5**	5	1.3313	0.6645	F52
1	3	2	4	**2**	5	**2.5**	4.5108	0.7619	F32
1	3	2	4	**6**	5	5	3.6338	1.1908	F9
1	3	**3**	4	**6**	5	5	4.1115	1.5028	F23
1	3	2	4	4	5	5	4.3247	0.888	F61

in between the input layer and the middle layer indicate the weight factor (IW {1, 1}) associated with input node, and bias input (b {1}) acts on a neuron like an offset. The blocks in between the middle layer and the output layer indicate the weight factor (LW {2, 1}) associated with hidden layer, and bias input (b {2}) acts on a neuron like an offset. Fig. 8 shows the performance plot for the trained neural network.

A. Simulation Results

The classifier results for a few randomly generated test patterns for the filter circuit are shown in Table V.

VI. FUZZY FAULT DIAGNOSIS SYSTEM

The first step in analysing the CUT is to study the response of the circuit to a stimulus signal that is exactly the same as that used for simulation of the circuit behaviour during the compilation of the fault dictionary. The stimulus signal, in this case, being a unit step signal, the output transient of the given circuit is recorded and the settling time and peak amplitude for that transient output signal are determined. These readings are then fed as inputs to the fuzzy inference system (FIS), which will carry out the task of generation of an output corresponding to the inputs given, the relationship being the inputs and outputs being a non-linear one.

Earlier approaches to the diagnosis of faults in analog circuits have relied upon the strength of probability theory to recognise patterns in the occurrence of faults assuming statistical random characteristics of the fault conditions and the surrounding environment. However, it seems a much simpler proposition to consider faulty networks as fuzzy systems since there is hardly a necessity to maintain exacting levels of accuracy while solving problems connected with the isolation of faults [15]. With the use of a fuzzy inference system (FIS), the potential for fuzzy logic in the development of a model that can characterize results through approximate reasoning is utilized for the purpose of identification of faults.

Fuzzy set theory is a discipline that revels in uncertainties and approximations rather than the precise and well-defined boundaries that can be seen associated with crisp sets, the property being termed as vagueness of fuzziness. As a consequence, members of a fuzzy set may possess partial membership to that set as opposed to the concept of crisp sets where an entity is either a member of the set, or it is not. This allows for flexible assignment of degrees of membership to entities based on their relationship to a set. Also, in fuzzy set theory, variables are described in terms of membership functions or "truth value" in relation to a particular fuzzy set, the values of the function lying in the range [0,1] [16].

Fuzzy inference systems are a direct application of fuzzy logic and fuzzy sets theory, otherwise known as fuzzy rule-based systems or fuzzy models. Fuzzy inference systems offer a great advantage in being compatible with linguistic concepts and rules that transpire naturally in human beings. In addition to this, they can be used to effortlessly map inputs and outputs that bear a non-linear relationship to each other [17]. A typical FIS is a combination of four sections namely:

- Fuzzifier
- Rule-base
- Inference engine
- Defuzzier

The fuzzifier transforms the crisp inputs given to the system into members of fuzzy sets, which are defined linguistically in agreement with human perception of the inputs. This step assigns degrees of truth to each input parameter. The fuzzy rule-base is usually made up of IF-THEN statements connected with logical operators, the main purpose being knowledge acquisition with regard to the input values. In the inference engine, the rules in the rule base are evaluated and an output fuzzy set is generated corresponding to each output variable. These output fuzzy sets are then converted into crisp values in the defuzzifier section.

The method of fault diagnosis under consideration has two time-domain specifications (settling time and peak amplitude) as the inputs to the FIS, from which a numerical value unique to each faulty configuration is derived. The numerical value, so obtained for each configuration, depends completely upon the membership functions assigned to each variable and the

rules that compose the fuzzy rule-base.

The Mamdani model has been chosen to construct the fuzzy inference system as it is very well suited to the simplification of problems involving linguistic rules based on human experience. The FIS proposed to solve the current problem of fault isolation takes the form shown in Fig. 9.
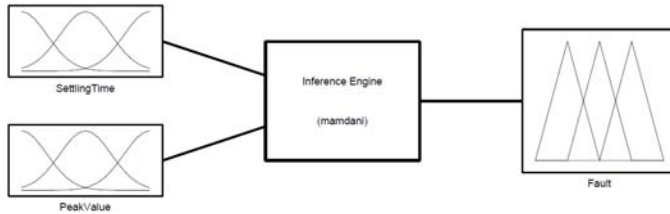
Fig. 9 Fuzzy inference system

The Mamdani fuzzy inference model has the distinct advantage of greater interpretability of quantities when compared to other fuzzy models such as the Takagi-Sugeno fuzzy logic system. Interpretability refers to the quality of possessing an inherent connection with the natural way of description of quantities through linguistic expressions. These linguistic variables are described by their respective membership functions, both the inputs and the outputs expected to be fuzzy variables.

The membership functions for the inputs to the current FIS, ie. the time-domain response parameters are as follows:
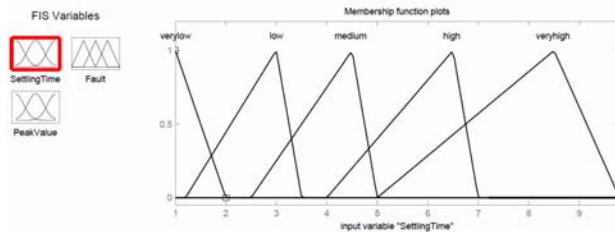
Fig. 10 Membership function for settling time

As seen above, the range of occurrence of each input variable is divided into five regions, namely-'Very low', 'Low', 'Medium', 'High', and 'Very high'.

The choice of areas over which these classifications
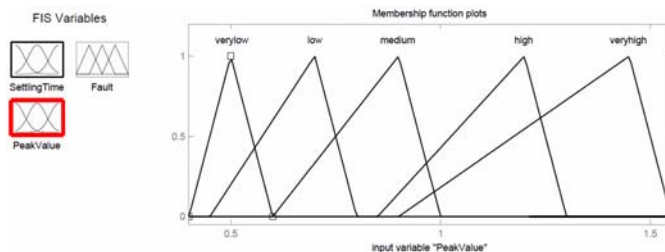
Fig. 11 Membership function for peak amplitude

extend depend solely upon the knowledge accumulated through simulation of the circuit for all the conditions, the

resulting effect being to reciprocate the designer's linguistic interpretation of the input values with relation to the range of possible values of the parameters. Such a representation of the time-domain specifications transforms them into linguistic parameters that can be input to the Mamdani FIS. The type of membership function shown in the Fig. 10 and 11 is known as *trimf* or 'triangular membership function', which is particularly efficient as far as computation time is concerned, due to its simple structure. In a similar fashion, the output variable is also assigned a membership function to make it a linguistic fuzzy variable.

The membership function shown in Fig. 12 is divided into areas that represent the different fault IDs that may be arrived at after processing of the inputs using the fuzzy rule base. The horizontal scale, containing values that will be displayed as the output fuzzy block, is chosen arbitrarily as per convenience, provided it allows for enough values to distinguish one faulty configuration from the other.
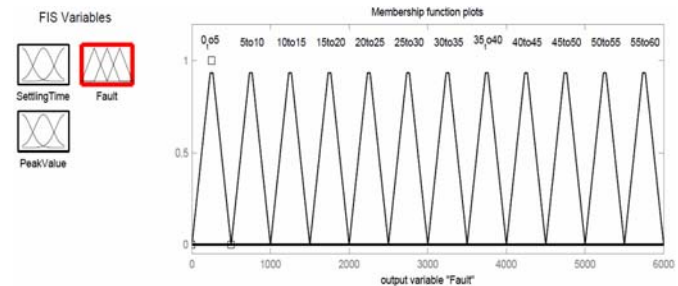
Fig. 12 Membership for fault ID

The fuzzy rule base is an aggregation of IF-THEN- rules that define the relationship between the input and output fuzzy sets. In effect, the use of linguistic variables and fuzzy IF-THEN- rules exploits the tolerance for imprecision and uncertainty. In this respect, fuzzy logic mimics the crucial ability of the human mind to summarize data and focus on decision-relevant information. Since there are 61 different configurations of the CUT in the fault dictionary, there must exist 61 fuzzy rules for the problem under consideration. The format of each of these rules is as follows:

IF
Settling Time is very low/low/medium/high/very high
OR
Peak Amplitude very low/low/medium/high/very high
THEN
Fault is 0-5/5-10/10-15...........50-55/55-60

OR denotes the fuzzy 'max' operator, which selects the maximum of two values. It is also used to symbolize the union of two fuzzy sets. After the OR operation between the two inputs, a fuzzy set is generated as the output. Finally, a defuzzification algorithm is implemented on the resultant output fuzzy set to calculate the output of the fuzzy system as a precise numerical value.

TABLE VII

FIS OUPUT FOR RANDOM FAULT CONDITIONS

Component Values (R in kΩ and C in nF)							FIS Inputs		FIS Output	Fault ID
R1	R2	R3	R4	R5	C1	C2	Settling Time	Peak Amplitude		
1	4.5	2	4	4	5	5	5.2859	0.8600	3113	F3
1	3	2	4	4	5	2.5	3.9499	1.1012	2977	F14
1	4.5	2	4	6	5	5	4.8269	1.1242	2872	F21
1	3	2	4	2	5	2.5	4.5108	0.7619	2912	F32
0.5	3	2	4	2	2.5	5	2.5534	0.6810	2874	F44
1.5	4.5	3	4	6	5	5	5.4507	1.1908	2792	F49
1	3	3	4	6	7.5	7.5	6.1672	1.5028	3223	F55
1	3	2	4	4	5	5	4.3247	0.888	2847	F61

The centroid method of defuzzification was employed on the fuzzy set 'Fault' to obtain a crisp value which can be identified easily to isolate faults. The centroid method, acclaimed to be the most reliable of all defuzzification methods, gives the value corresponding to the centre of the area of the fuzzy set as the crisp output, favouring the rule with the greatest output area. The crisp value so obtained as the output of the FIS is unique to each faulty configuration, a property that can be exploited for the identification of faults in components that constitute the circuit under test.

The formulation of the rules for the current problem of fault diagnosis can be clearly portrayed through the use of a Fuzzy Associative Memory (FAM) table. Fuzzy associative memory (FAM) is a rule-based system based on fuzzy sets and logic. Fuzzy associative memories embody a bank of fuzzy rules that reflect expert knowledge in linguistic form. A compound FAM rule is a compound linguistic condition: "If X1 is A1 and X2 is A2. …and Xn is An then Y is B". A fuzzy associative memory can be used to combine associative memory and fuzzy logic, thus encoding the fuzzy output set with the fuzzy input set. The framework of FAM allows us to adaptively add and modify fuzzy rules, directly from experts or from statistical techniques. A fuzzy associative matrix is constructed to map the input to the associated output by a max–min composition operation. If an input is fuzzy (a degree of membership is provided to indicate the closeness), the output will also be a fuzzy set. FAM is transparent in the sense that knowledge is explicitly revealed in the rules. Here, the two fuzzy variables, settling time and peak amplitude can be laid out in a 2D matrix such that one variable represents each axis. Each entry in the matrix is the output corresponding to a logical proposition defined by the values of the variables in that particular row and column, as shown in Table VI.

The outputs of the suggested FIS for a few randomly manipulated faults are listed in Table VII. It can be inferred from the table that each faulty configuration gives a unique FIS output thus reducing the process of fault identification to that of mere numerical matching with the previously created fault dictionary for the same circuit.

VII. DISCUSSION

Although both fuzzy and neural systems are equally accurate in identifying the faulty configurations, an analysis of the paths taken to reach that final step yields sufficient parameters to compare the two systems. Firstly, the time taken to complete the process is longer with neural systems as compared with the time taken by fuzzy systems. An increase in the size of the circuit noticeably increases the time taken by neural networks, this factor not playing a major role in the case of fuzzy systems.

TABLE VI

FUZZY ASSOCIATIVE MEMORY TABLE

Peak Amplitude \ Settling Time	Very Low	Low	Medium	High	Very High
Very Low	F25-F30 F45-F50 F55-F60	F40-F45 F45-F50 F50-F55	F15-F20 F35-F40 F55-F60	F45-F50 F50-F55	--------
Low	F5-F10 F20-F25 F25-F30	F10-F15 F20-F25 F25-F30 F40-F45	F0-F5 F15-F20	F5-F10 F15-F20	F5-F10 F35-F40
Medium	--------	F0-F5 F5-F10 F10-F15 F30-F35	F0-F5 F15-F20 F25-F30 F30-F35 F55-F60	F0-F5 F10-F15 F20-F25 F40-F45 F45-F50	F20-F25 F25-F30
High	F35-F40 F50-F55	F40-F45	F10-F15 F15-F20 F30-F35	F20-F25 F40-F45 F45-F50	F50-F55
Very High	--------	F30-F35	F35-F40 F55-F60	F50-F55	--------

This difference in times taken for processing can be attributed to the computational burden that is laid on the processor(s) by either system. Neural networks, with multiple hidden layers, tend to require greater processing in comparison to fuzzy systems. Even with fuzzy systems, experimentation was done with two membership functions - the triangular membership function (*trimf*) and the Bell membership function (*bellmf*). The results obtained using both these

functions were equally good, there being no major difference in their ability to distinguish between the various fault conditions. However, it is more favourable to use the triangular membership function as it is computationally lighter than the Bell membership function by virtue of its simple structure.

Apart from this, fuzzy systems can be expected to be more robust than neural systems because of the 'fuzziness factor' that allows deviations from the originally programmed pattern. Finally, choice of system must be made keeping in mind the kind of data that is available for use. Data that is readily available in structured format is ideally suited for use with fuzzy systems. On the other hand, in circumstances where numerical data is available but there is no apparent way to structure the data, neural systems are more convenient. The structuring can arise out of human perception of quantities, a defining quality of fuzzy logic systems. However, prior knowledge concerning the data is a necessary requirement in the design of fuzzy systems. Neural networks do not require prior knowledge as they are capable of learning and evolving through a number of learning algorithms.

VIII. CONCLUDING REMARKS

The above discussion carries weight mainly when there are constraints on the memory capacities and operating frequencies of the processors used to carry out fault diagnosis. In the presence of high performance processors, the disadvantages of prolonged processing duration in neural networks are effectively cancelled out, thus equating the efficiencies of both fuzzy and neural systems. Hence, depending upon the availability of hardware, one must decide upon the priority to be given to reduction in processing time in order to be better able to choose between neural and fuzzy logic systems. Further, for certain applications where robustness of the system is awarded more importance, a fuzzy system will be more beneficial. The experience of the programmer in dealing with datasets similar to the one in consideration also influences a choice in favour of fuzzy logic. On the other hand, a lack of such familiarity automatically tilts the balance in favour of neural networks. Hence, based on these parameters, a judicious choice of system should be arrived upon.

REFERENCES

[1]   J. W. Bandler and A. E. Salama, "Fault diagnosis of analog circuits," Proc. IEEE, vol. 73, 1985, pp. 1279–1325.

[2]   J. A. Starzy and J. W. Bandler, "Multiport approach to multiple fault location in analog circuits," IEEE Trans. Circuits Syst., vol. 30, 1983, pp. 762–765.

[3]   M. Tadeusiewic and M. Korzybski, "A method for fault diagnosis in linear electronic circuits," Int. J. Circuits Theory Applications, vol. 28, 2000, pp.254–262.

[4]   G. Fedi, R. Giomi, A. Luchetta, S. Manetti, and M. C. Piccirilli, "On the application of symbolic techniques to the multiple fault location in low testability analog circuits," IEEE Trans. Circuits Syst. II, vol. 45, Oct. 1998, pp.1383–1388.

[5]   R. Spina and S. Upadhyaya, "Linear circuit fault diagnosis using neuromorphic analyzers," IEEE Trans. Circuits Syst. II, vol. 44, Mar. 1997, pp. 188–196.

[6]   M. Catelani and M. Gori, "On the application of neural network to fault diagnosis of electronic analog circuits," Measurement, vol. 17, 1996, pp.73–80.

[7]   W. Hochwald and J. D. Bastian, "A dc approach for analog fault dictionary determination," IEEE Trans. Circuits Syst. I, vol. 26, May 1979 pp. 523–529.

[8]   K. C. Varghese, J. H.Williams, and D. R. Towill, "Simplified ATPG and analog fault location via a clustering and separability technique," IEEE Trans. Circuits Syst., vol. 26, May 1979, pp. 496–505.

[9]   A. Mckeon and A. Wakeling, "Fault diagnosis in analogue circuit using AI technique," in IEEE Int. Test Conf., 1989, pp. 118–123.

[10] Gertler,J.,J.,.fault Detection and Diagnosis in Engineering Systems",. *Marcel Dekker*.New York, 1998.

[11] Chen,J.,Patton,R.J. "Robust Model-Based Fault Diagnosis for Dynamic systems",.*Kluwer Academic Publishers*. Masssachusetts, 1999.

[12] V. Manikandan and N. Devarajan **"**SBT Approach towards Analog Electronic Circuit Fault Diagnosis" *Hindawi Publishing Corporation, Active and Passive Electronic Components,* Volume 2007, Article ID 59856, 11 pages.

[13] L.Fausrtt, "Fundamentals of Neural Networks," Prentice-Hall, Upper Saddle River, NJ, USA, 1994.

[14] S.HSU, et al., **"**Comparative analysis of five neural network models," Remote Sensing Review, Vol.6, 1992, pp.319-329.

[15] Jonghee Lee and Samuel D. Bedrosian, "Fault isolation algorithm for analog electronic systems using the fuzzy concept", IEEE Transactions on Circuits and Systems, Vol. CAS-26, No. 7, July 1979.

[16] Dubois, Prade, "Fuzzy sets and systems," Academic Press, New York, 1980.

[17] Serge G, "Designing Fuzzy Inference Systems from Data: Interpretability oriented Review", IEEE Transactions on Fuzzy Systems, 2001.

AUTHORS PROFILE

V.Prasannamoorthy is currently Senior Grade Lecturer in the Department of Electrical Engineering, Government College of Technology, Coimbatore. (phone: +919443750031, e-mail: prasanna_gct1995@yahoo.com)

R. Bharat Ram is pursuing B.E. degree in Electrical and Electronics Engineering at Government College of Technology, Coimbatore.

V. Manikandan is currently Assistant Professor in the Department of Electrical Engineering, Coimbatore Institute of Technology, Coimbatore.

N.Devarajan is currently Assistant Professor in the Department of Electrical Engineering, Government College of Technology, Coimbatore.

Evaluation of English-Telugu and English-Tamil Cross Language Information Retrieval System using Dictionary Based Query Translation Method

P. Sujatha
Department of Computer Science
Pondicherry Central University
Pondicherry-605014, India.
spothula@gmail.com

P. Dhavachelvan
Department of Computer Science
Pondicherry Central University
Pondicherry-605014, India.
dhavachelvan@gmail.com

V.Narasimhulu
Department of Computer Science
Pondicherry Central University
Pondicherry-605014, India.
narasimhavasi@gmail.com

Abstract—**Cross Lingual Information Retrieval (CLIR) system helps the users to pose the query in one language and retrieve the documents in another language. We developed a CLIR system in computer science domain to retrieve the documents in Telugu and Tamil languages for the given English query. We opted for the method of translating queries for English-Tamil and English-Telugu language pairs using bilingual dictionaries. Transliteration is also performed for the named entities present in the query. Finally, the translation and transliteration results are combined and used the resultant query to the searching module for retrieving target language documents. For Telugu, we achieve a Mean Average Precision (MAP) of 0.3835 and for Tamil, we achieve a MAP of 0.3665.**

Keywords-Cross Lingual Information Retrieval; Translation; Transliteration; Ranking.

I. INTRODUCTION

CLIR can be defined as a subfield of Information Retrieval (IR) system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query. It Facilitates the process of finding relevant documents written in one natural language with automated systems that can accept queries expressed in other language(s) is thus the major purpose of CLIR system. The process is bilingual when dealing with a language pair, that is, one source language and one target or document language. In multilingual information retrieval the target collection is multilingual, and topics are expressed in one language [1]. In any of such cases CLIR is expected to support queries in one language with a collection in another language(s) [2].

According to Peters and Sheridan [3] CLIR is a complex multidisciplinary research area in which methodologies and tools developed in the field of IR and natural language processing converges. IR is traditionally based on matching the words of a query with the words of document collections. Because the query and the document collection are in different languages, this kind of direct matching is impossible in CLIR. Translation is needed: either the query has to be translated into the language of the documents or the documents have to be translated into the language of the query. Obviously, translating the whole document collection is more demanding, as it requires more scarce resources like full-fledged Machine Translation (MT) system, which is not available for a number of languages in developing countries. Hence query translation techniques become more feasible and common in development and implementation of CLIR system. The present paper discusses a CLIR system using query based translation.

The organization of the paper is as follows. Section II, describes related work done on CLIR systems in Indian languages. Section III discusses a brief overview of CLIR system architecture. Evaluation results are described in Section IV. The conclusion and future enhancements of the paper are given in Section V.

II. RELATED WORK

Many organizations in India are working on the CLIR system for different Indian Languages [13]. IIIT, Hyderabad has developed a Hindi and Telugu to English CLIR system [4]. They used a vector based ranking model with bilingual lexicon using word translations combined with a set of heuristics for query refinement after translation. Jagadeesh and Kumaran [5] build a CLIR system with the help of a word alignment table learned from a parallel corpus, primarily for statistical machine translation. They participated in the Cross Language Evaluation Forum (CLEF) competition, in the Indian language sub-task of the main Ad-Hoc monolingual and bilingual track. This track tests the performance of systems in retrieving the relevant documents in response to a query in the same and different languages from that of the document set. In Indian context, documents are provided in English (corpus) and queries are specified in different languages including Hindi, Telugu, Bengali, Marathi and Tamil on the CLEF dataset. A cross-language query focused

multi-document summarization for the Telugu-English language pair was described in [6]. The authors used a cross-lingual relevance based language modeling approach to generate extraction based summary. It would provide a syntactically well formed set of sentences in the summary to enable easy machine translation. Other benefit of the system is output can be an easily translatable content (minimizing ambiguities). Mandal et al. [7] described two cross-lingual and one monolingual English text retrieved at CLEF in the Ad-Hoc track. The cross-language task includes the retrieval of English documents in response to queries in two most widely spoken Indian languages Hindi and Bengali. Here, authors adapted automatic query generation and machine translation approach to develop the system.

An Indian Language Information Retrieval System [8], which exploits the significant overlap in vocabulary across the Indian languages. Cognates are identified using some of the well-known similarity measures, and incorporated this with the traditional bilingual dictionary approach. The effectiveness of the retrieval system was compared on various models. The results show that using cognates with the existing dictionary approach leads to a significant increase in the performance of the system. Language independent information retrieval is one of the major issues in the web access by the regional population of any kind. Language Independent Information Retrieval from Web (LIIRW) was described in [9]. Here, the user with the independence of typing the query in any language of his choice and getting the results in any language or any combination of languages, it is intended to make the multilingual content of the web easily available and more noticeable. It addresses the implementation of the LIIRW concept in Indian languages (Hindi and Tamil). A Tamil-English CLIR system was developed in [10]. This system is mainly developed for the farmers of Tamilnadu in Agriculture domain. It helps them to specify their information need in Tamil and retrieve the documents in English (corpus). Here, the query in Tamil language is translated syntactically and semantically to English using statistical machine translation approach and gives the better result. The system exhibits a dynamic learning approach.

This paper presents a CLIR system, which translate English query into Tamil and English query into Telugu using bilingual dictionaries related to computer domain. It also transliterates the named entities, which are present in the query other than the words which can be translated.

III. System Architecture

The overview of CLIR system is shown in Fig. 1. It mainly contains the following modules: Text Processing, Verification, Translation, Transliteration and Retrieval and Ranking.

Text Processing: For a given source or user query, this module performs preprocessing of the source query. That is, before translating the source query into target query, need to perform some text processing steps. The following are the text processing steps: Tokenization, stop words removal, morphological analyzer and stemming. Tokenization is the task of dividing query into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. These tokens are referred to as terms or words. After tokenization, some words in the query need to be removed called stop words, which should not help retrieving target documents. Examples of those words are: *the, is, was, that etc*. Using stop words removal step, these words can be removed from the source query. Morphological analyzer analyzes the structure of the words in the query. Examples of those words are *verbs, adverbs, adjectives etc*. That is, vocabulary of the words in the query can be identified. Stemming is the process of reducing inflected words to their base or root form. For example, *fishing, fished and fisher* are inflected words, which can be reduced into their root form *fish* using stemmer. After the text processing, the output of the source query *(SQ)* is called preprocessed source language query *(PSQ)*, which includes preprocessed source language query words $\{PSW_1, PSW_2...PSW_n\}$.

Verification Module: It is designed for the purpose of checking the occurrence of source language words in Machine Readable Dictionaries (MRD) or Bilingual (source to target language) dictionaries. MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. After text processing module, the verification module accepts the input query words $\{PSW_1, PSW_2...PSW_n\}$ and performs a database lookup operation to check whether the given query is directly present in the bilingual dictionary. The words which found in the dictionary $\{PSW_1, PSW_2...PSW_i\}$ can be given to the Translation Module and the words which are not found in the dictionary $\{PSW_1, PSW_2...PSW_j\}$ can be given to the Transliteration module.

Translation Module: In this paper, the language of the user query is English and the documents considered for retrieval are in Tamil and Telugu languages. These documents are a set articles based on computer science terminology. Hence we have concentrated on queries with computer terminology. It is also called machine translation module. It follows the dictionary based translation method. Dictionary based translation method can translates the query words using the bilingual dictionaries. These words are called vocabulary words. We have developed an English-Tamil and English-Telugu bilingual dictionaries that contain most the words related to computer science domain. The dictionary had to be built from the scratch as no resource is available for this domain. After each intermediary step in the Morphological Analyzer, the extracted word is mapped with the bilingual dictionary to check whether it is a root word. If it is available, meaning of the word is returned. If not, the word is then passed on to the subsequent stages in the Morphological Analyzer. The words which are not found in the dictionary are called Out Of Vocabulary (OOV) words. This module takes input as $\{PSW_1, PSW_2...PSW_i\}$ and translates into target

language query words *{TDW₁, TDW₂…TDWᵢ}* using bilingual dictionary. The output of this module is translated dictionary words $\{TDW_1, TDW_2...TDW_i\}$.

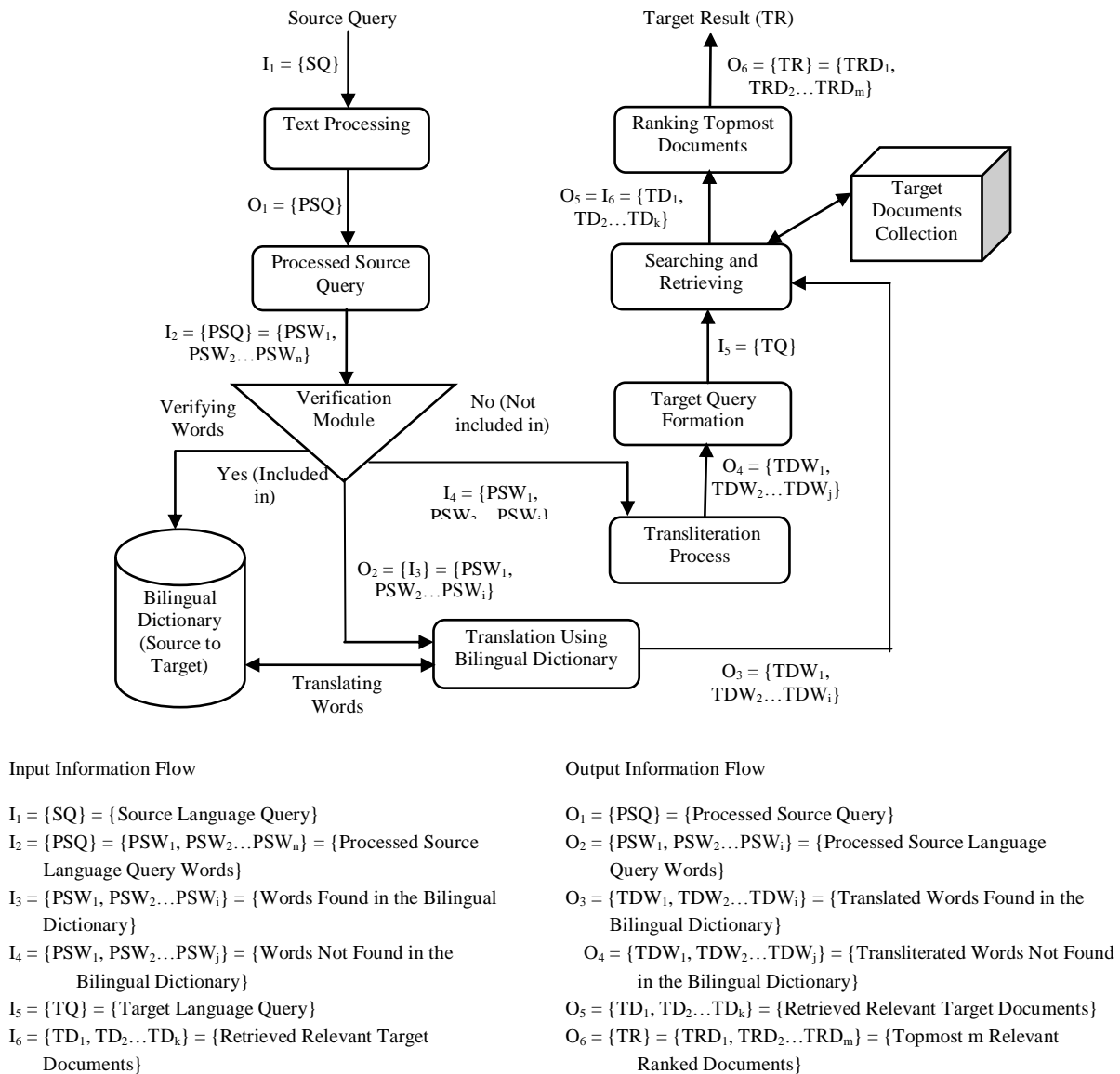source language into a target language without the aid of a resource like a bilingual dictionary.

Figure 1. System Architecture

Input Information Flow

$I_1 = \{SQ\} = \{$Source Language Query$\}$

$I_2 = \{PSQ\} = \{PSW_1, PSW_2...PSW_n\} = \{$Processed Source Language Query Words$\}$

$I_3 = \{PSW_1, PSW_2...PSW_i\} = \{$Words Found in the Bilingual Dictionary$\}$

$I_4 = \{PSW_1, PSW_2...PSW_j\} = \{$Words Not Found in the Bilingual Dictionary$\}$

$I_5 = \{TQ\} = \{$Target Language Query$\}$

$I_6 = \{TD_1, TD_2...TD_k\} = \{$Retrieved Relevant Target Documents$\}$

Output Information Flow

$O_1 = \{PSQ\} = \{$Processed Source Query$\}$

$O_2 = \{PSW_1, PSW_2...PSW_i\} = \{$Processed Source Language Query Words$\}$

$O_3 = \{TDW_1, TDW_2...TDW_i\} = \{$Translated Words Found in the Bilingual Dictionary$\}$

$O_4 = \{TDW_1, TDW_2...TDW_j\} = \{$Transliterated Words Not Found in the Bilingual Dictionary$\}$

$O_5 = \{TD_1, TD_2...TD_k\} = \{$Retrieved Relevant Target Documents$\}$

$O_6 = \{TR\} = \{TRD_1, TRD_2...TRD_m\} = \{$Topmost m Relevant Ranked Documents$\}$

Transliteration Module: Translation module can handle only vocabulary words, but not OOV words. Previous studies suggested that OOV words can be properly handled; otherwise the retrieval performance of CLIR system can reduce up to 60% [11]. OOV terms can be of many types. They can be of newly formed words, loan words; abbreviations or domain specific terms etc. One possible and effective way of handling OOV terms is using transliteration techniques. Transliteration is the suitable method for translating OOV terms. Transliteration is the process of transforming a word in a

This work follows grapheme based transliteration model [12], which is one of the major techniques of transliteration. Grapheme refers to the basic unit of written language or smallest contrastive units. In grapheme based transliteration model spelling of the original string is considered as a basis for transliteration. It is referred to as the direct method because it directly transforms source language graphemes into target language graphemes without any phonetic knowledge of the source language words. This module takes input as *{PSW₁, PSW₂...PSWⱼ}*. It is designed with following steps: dividing

each word into characters, character level alignment, mapping with intermediate target scheme and generation of each target word. First it divides each processed word which is not found in the dictionary *{PSW₁, PSW₂...PSWⱼ}* into individual characters. Next it applies character level alignments on *{PSW₁, PSW₂...PSWⱼ}* which include level wise alignments. After alignment, it uses intermediate mapping (Roman) scheme, which is suitable for current character level alignments for mapping target language characters. Using intermediate scheme, it maps individual characters in the source word with the individual target characters and generates complete target word. Each target word can be generated in similar fashion. The output of this module is transliterated target words not found in the dictionary *{TDW₁, TDW₂...TDWⱼ}*. For example, the transliteration of an English word (data mining) into Telugu is shown in Table I. The transliteration of an English word (program) to Tamil is also shown in Table II. Here, English word is scanned from left to right and divided according to character level alignments based on the target languages (Telugu or Tamil). The final Telugu or Tamil word is generated based on the Romanization scheme. That is, "data mining" is transliterated into "డేట మైనింగ్" in Telugu and

"program" is transliterated into "ப்ரோக்ராம்" in Tamil.

TABLE I. ENGLISH TO TELUGU TRANSLITERATION EXAMPLE

English word	Transliteration in Telugu
Da	డే
Ta	డేట
Mi	డేట మై
Ni	డేట మైని
N	డేట మైనిన్
G	డేట మైనింగ్

TABLE II. ENGLISH TO TAMIL TRANSLITERATION EXAMPLE

English word	Transliteration in Tamil
p	ப்
ro	ப்ரோ
g	ப்ரோக்
ra	ப்ரோக்ர
m	ப்ரோக்ரம்

Retrieval and Ranking Module: This module is designed for searching and retrieving relevant target documents for a

given query. Both translated words {*TDW₁, TDW₂...TDWᵢ*} and transliterated words *{TDW₁, TDW₂...*TDWⱼ} combined forms the target language query *(TQ)*. For merging these translated and transliterated words, we have used a simple array-based technique. That is, each word in the query will be numbered. Once their translation and transliteration tasks are completed arrange the resultant words in the original order. Using *TQ*, target language documents can be retrieved. Target language documents are collected from online and stored in the database. For a given TQ, search process will retrieve the relevant target documents *{TD₁, TD₂...TDₖ}* from the target documents collection. Here, search process is designed with indexing method. Indexing is the simple and fast method for retrieving relevant documents. Retrieved documents are given to the ranking method [13] for making final ranking which is described as below.

Given a pair of cross lingual queries *(qₑ, qₜₑ)* and *(qₑ, qₜₐ)*, we can extract the set of corresponding cross lingual document pairs and their click counts *{(eᵢ, teⱼ), (C(eᵢ),C(teⱼ))} and {(eᵢ, taⱼ), (C(eᵢ),C(taⱼ))}*, where $i = 1, \ldots, N$ and $j = 1, \ldots, n$. Based on that, we produce a set of cross lingual ranking instances *S = {φᵢⱼ, zᵢⱼ}*, where each φᵢⱼ = {xᵢ; yⱼ; sᵢⱼ} is the feature vector of *(eᵢ, teⱼ)* and *(eᵢ, taⱼ)* consisting of three components: xᵢ = f*(qₑ, teᵢ)* and f(qₑ, taᵢ) is the vector of monolingual relevancy features of eᵢ, yᵢ = f*(qₜₑ, teⱼ) and f(qₜₐ, taⱼ)* is the vector of monolingual relevancy features of teⱼ and taⱼ, and sᵢⱼ = *sim(eᵢ, teⱼ) and sim(eⱼ, taⱼ)* is the vector of cross-lingual similarities between eᵢ and teⱼ and eᵢ and taⱼ, and zᵢⱼ = (C(eᵢ),C(teⱼ)) and (C(eᵢ),C(taⱼ)) is the corresponding click counts. The task is to select the optimal function that minimizes a given loss with respect to the order of ranked cross lingual document pairs. For each pair of cross lingual queries *(qₑ, qₜₑ)* and *(qₑ, qₜₐ)*, the documents were ranked using Lucene's BM25[1] algorithm as the similarity metric.

This ranking method specifies topmost m relevant documents *{TRD₁, TRD₂...TRDₘ}* from a given set of k documents *{TD₁, TD₂...TDₖ}*. *{TRD₁, TRD₂...TRDₘ}* are shown to the user as a final target result.

IV. EVALUATION RESULTS

We have used two cross lingual runs: E->Te and E->Ta. In this paper, the official English topics are used to retrieve Telugu and Tamil documents. The English topics are translated into Telugu and Tamil by the model prescribed in Fig. 1. Target documents contain collection of both Tamil and Telugu language documents. The details about the number of target document collection are given in Table III. Most of the documents (Tamil and Telugu) are collected from the electronic news articles. These documents covered only computer science articles. Remaining documents are collected from the native language websites for a period of three to four months. The details about the total number of terms, number of unique terms and average document length is specified in Table III.

---

1   http://nlp.uned.es/~jperezi/Lucene-BM25/

TABLE III. DETAILS OF DOCUMENTS COLLECTION

Number of documents	18659
Number of terms	5848964
Number of unique terms	84594
Average document length	85

The test set for evaluating the performance consists of 100 English queries. 50 queries on retrieving Telugu language documents and remaining 50 queries are used on retrieving Tamil language documents. Finally, either language documents are ranked based on the relevance with the specified query. The performance results are measured in terms of metrics like Recall, MAP, P@10, R-Precision as shown in Table IV. The performance result shows that the given CLIR system retrieves the relevant documents in either language for a given query in English.

TABLE IV. PERFORMANCE EVALUATION RESULTS E-Te AND E-Ta EXPERIMENTS

Cross lingual runs	Recall	MAP	P@10	R-Precision
E -Te	86%	0.3835	0.4631	0.3820
E -Ta	84%	0.3665	0.4270	0.3647

The performance curves of E-Te and E-Ta runs are depicted in the Fig. 2 and Fig. 3 respectively. There is little difference in these two runs, i.e. the E-Te run outperforms the E-Ta run because the words in E-Te dictionary are better than the words in E-Ta dictionary i.e the quality of the dictionary will affect the performance of the system. For Telugu, we achieve a MAP of 0.3835 and for Tamil; we achieve a MAP of 0.3665. The recall levels in Telugu are 86%. The recall levels in Tamil are 84%.
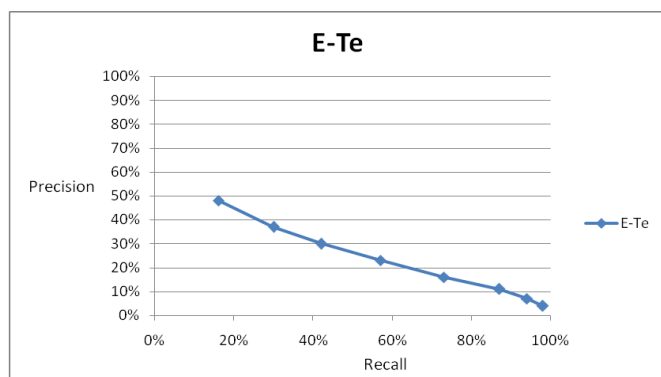
Figure 2. Average Precision vs Recall for E-Te run
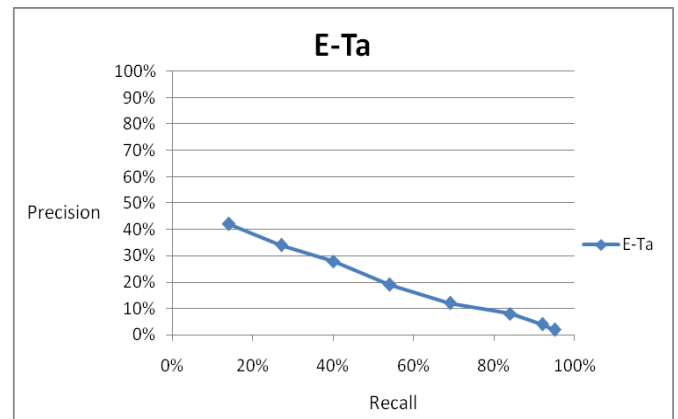
Figure 3. Average Precision vs Recall for E-Ta run

V. CONCLUSION AND FUTURE WORK

A CLIR system is developed for computer Science domain. The system focuses on dictionary-based approach that has been used for E-Te and E-Ta translation. Transliteration is also done using simple grapheme based transliteration model. In future we would compare the performance of this query translation method with document translation method. This system can be further extended to exhibit a dynamic learning approach wherein any new word that is encountered in the transliteration process could be updated in the database by allowing the user dynamically to insert it into the database along with its corresponding Tamil or Telugu transliterated words.

REFERENCES

[1] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and K. Jrvelin, "Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002," In Information Retrieval, 2004.

[2] E. Waterhouse, "Building translation lexicons for proper names from the web," In Thesis, Department of Computer Science, University of Sheffield, 2003.

[3] C. Peters and P. Sheridan, "Multilingual information access," In ESSIR '00: Proceedings of the Third European Summer-School on Lectures on Information Retrieval-Revised Lectures, pages 51–80, London, UK, 2001. Springer-Verlag.

[4] P. Pingali and V. Varma, "Hindi and Telugu to English cross language information retrieval at CLEF 2006," In Working Notes for the CLEF 2006 Workshop (Cross Language Adhoc Task), 20-22 September, Alicante, Spain.

[5] J. Jagadeesh and K. Kumaran, "Cross-lingual information retrieval System for Indian Languages", Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, pages: 80-87.

[6] P. Pingali and V. Varma, "Experiments in cross language query focused multi-document summarizations", Workshop on Cross Language Information Access CLIA-2007, International Joint Conference on Artificial Intelligence (IJCAI), 2007.

[7] D. Mandal, S. Dandapat, M. Gupta, P. Banerjee, and S. Sarkar, "Bengali and Hindi to English cross-language Text Retrieval under Limited Resources", In the working notes of CLEF 2007.

[8] M. Ranbeer, P. Nikita, P. Prasad, and V. Vasudeva, "Experiments in Cross-Lingual IR among Indian languages", International Workshop on Cross Language Information Processing (CLIP-2007), 2007.

[9] R. Seethalaksmi, A. Ankur, and R. Ranjit, "Language independent information retrieval from web", ULIB Conference 2007.

[10] D. Thenmozhi and C. Aravindan, "Tamil-English cross lingual information retrieval system for agriculture society", International Forum for Information Technology in Tamil (INFITT), Tamil International Conference 2009.

[11] D. Demner-Fushman and D. W. Oard, "The effect of bilingual term list size on dictionary-based cross-language information retrieval," In 36th Annual Hawaii International Conference on System Science HICSS'03), pp. 108-118, 2003.

[12] P. Majumder, M. M. Swapan parui, and P. Bhattacharyya, "Initiative for Indian Language IR Evaluation," Invited paper in EVIA 2007 Online Proceedings.

[13] W. Gao, J. Blitzer, M. Zhou, and K. F. Wong, "Exploiting Bilingual Information to Improve Web Search," Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1075–1083,Suntec, Singapore, 2-7 August 2009.

[14] J. H. Oh and K. S. Choi, "An ensemble of transliteration models for information retrieval," Information Processing and Management: an International Journal, v.42 n.4, pp. 980-1002, July 2006.

A Novel Approach for Hand Analysis Using Image Processing Techniques

Vishwaratana Nigam, Divakar Yadav, Manish K Thakur

Department of Computer Science & Engineering and Information Technology
Jaypee Institute of Information Technology, Noida, India
vishwaratna.nigam@gmail.com, divakar.yadav0@gmail.com, mthakur.jiit@gmail.com

Abstract- **Palmistry is the art of characterization and foretelling the future through the study of the palm, also known as palm reading, or chorology. With the help of palm lines and fingers one can know the characteristics as well as can foretell the future of a person but still this field is not much technically developed and a person has to analyze hands personally. In this paper we propose a ratio based system to characterize persons on the basis of their palm width-length and their finger length. We applied image processing techniques to generate and analyze the results.**

Keywords- Palmistry, Palm-width (Pw), Palm-length (Pl), Finger-length (Fl), Jupiter ruled, Saturn ruled, Sun ruled, Mercury ruled.

I. Introduction:

Palmistry is popular since ancient age. There are examples of stone-age showing human interest in palmistry. Emperor of china used hand and thumb prints to seal documents. In our Vedas also there is information available on palm analysis. Palmistry is the knowledge of reading and analyzing palm lines, fingers and getting results [6].

This field is not yet technically rich. With the help of this paper we are trying to make a system intelligent enough to generate the results. The system comprises of three basic components- existing palmistry knowledge, derived algorithms and image processing.

Human palms have some common characteristics in the form of mountains known as planets. The finger's length corresponding to the planet represents the strength of that planet for that person. Every planet has its own predefined qualities and when a person is ruled by the planet he inherits the qualities of the planet. Apart from planets, person's palm width-length, palm-length ratio and finger-length also tells about person's qualities and his nature.

So, with the help of palmistry one can know somebody's personality type, his characteristics, hidden skills, fortune and about the profession suits best to him. Till now the palmistry is a practical knowledge and we have no systemized approach to read hands. We don't have a system which can read human

hand and characterize them on the basis of their finger length - palm length and their ratios.

With the help of image processing and ratio based algorithms mentioned below, one can characterize people and may know their personality type as well.

1. Palm-width and Palm-length ratio Algorithm
2. Palm-length and Finger-length Ratio Algorithm
3. Finger length ratio Algorithm

The paper is organized as follows:
In the current section we have given the introduction to the basics of palmistry and the challenges which are faced while analysing hands. In section 2 we discuss about the literatures related to palmistry and image processing. In section 3 we propose a new approach to analyze palms and get results based on palmistry. Section 4 consists of experimental results performed. Finally conclusion is being given in section 5 followed by references.

II. Fundamentals

The system proposed in this work is novel as we are using traditional knowledge of palmistry and making it systemized and computer readable. Earlier researchers have done work in getting edges from pictures and we are taking help from these works and applying our ratio-system.

In this work we are trying to integrate palmistry knowledge with Image processing. So, we have used image processing knowledge to extract lines, then palmistry knowledge to get results. For image processing we are trying mainly line detection and curve detection. We found that for line and curve detection Hough Transform is most efficient algorithm. Hough transform makes an array and works as a matrix, so its functioning is easy. With the help of Hough transform we extract outlines of hand and lines on palm from the given hand print and then system will analyze those lines with the lines in our database [1], [2].

In order to help Hough-Algorithm to work better, we first apply *Canny-Algorithm.* It detects edges accurately so, it makes things easier for Hough-algorithm [3], [4].Then next step followed is pattern matching for getting boundaries of palm [5]. Pattern of input palm image is matched with the palm image templates maintained in the database pixel by pixel and by taking difference between both images (pixel by pixel) we get one out of three possible results: less match, more match and exact match. The third possible result i.e. exact match is practically impossible and the first possible result i.e. less match is not suitable for our method, so we consider the second possible result i.e. more match. The second possible result is suitable for our system as we are only concerned for pattern of palm that gives the idea of boundaries of palm. Once we get boundaries of palm we apply pixel addition and difference methods to compute pixel distances [11] and thereof palm width (Pw), palm length (Pl), and finger length (Fl). Using Pw, Pl, and Fl we compute the ratios, later used in the proposed ratios based system discussed in next section.

In palmistry field some palmists have given their logics about fingers' length, palm length-width and its effect on person's personality, as every person has some unique characteristics and different nature and that is ruled by his hand, his hand-type, finger length and ruling planet [6], [10]. So, by integrating that existing knowledge we are proposing our ratio-system, which will perform a ratio-based analysis on palm and fingers and generate a result based on palmistry[7], [8].

III. PROPOSED SOLUTION APPROACH

The proposed system for hand analysis works in two separate approaches. The first approach is ratio based system and second approach is based on finger length comparison.

A. Ratios Based System Approach

With the help of ratio based system characterization of personalities may be achieved. The proposed system works in two steps.

Step 1: Compute the Palm-width and Palm-length ratios i.e. Pw/Pl as shown in Fig. 1. Based on the ratios computed, the palm images are categorised in either of two categories: Square or rectangular using following conditions.

Case 1: If (Pw/Pl) > 0.8 then square palm
Case 2: If (Pw/Pl) < 0.8 Rectangular palm
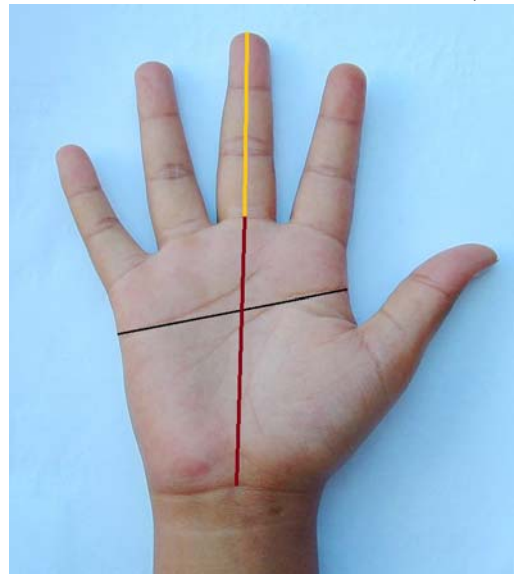
Fig. 1. Palm Image showing width- length and finger length

Step 2: Once the input palm images are characterized in square or in rectangular palm category, then they are further subdivide into four sub-categories based on the ratios between palm-length (Pl) and Finger-length(Fl) i.e. Pl/Fl as follows:

    I.   Square palm + Short fingers
    II.   Square palm + Long fingers
    III.   Rectangular palm + Short fingers
    IV.   Rectangular Palm + Long Fingers

The divisions of palm images into above four sub-categories are based on following conditions:

Case1: If (Fl/Pl) < 0.9   then Square hand +Short fingers
Case2: If (Fl/Pl) > 0.9   then Square hand + long fingers
Case3: If (Fl/Pl) < 0.8 then Rectangular Hand + Short fingers
Case4: If (Fl/Pl) > 0.8   then Rectangular hand + Long fingers

Based on above sub-categorization following analysis is performed and thereof positive as well as negative characteristics are drawn for a given palm-image.

1.   Square palm + short fingers

   Positives:- Reliability, orderliness, tolerance, constructive attitude.
   Negatives: - Insensitive, materialistic, overcautious, dislike changes.

2.   Square palm + long fingers

   Positives: - independent, self starters, political mind.
   Negatives:- Dislikes superiors, find difficult to work under another's flag.

3.   Rectangular + Short fingers-

Positives:- High energy, work best under pressure, enthusiastic, expansive, work well in short term goals.

Negatives: - Destructive behavior, cruel, self centered, don't like criticism, need deadlines.

4. Rectangular + Long fingers-

Positives:- Sensitive, intuitive, compassionate, do well in sales and public relations.
Negatives: - Less friendly, depressive, amoral.

So, with the help of this ratio system one can characterize the human palm in above four sub-categories. Now we will apply our ratio system on fingers and get different personality types of different people on the basis their finger lengths.

B. Finger Length Comparison Based Approach

This approach is based on the length of finger's partitions. As in normal cases every finger has three partitions. The length of each partition for every finger is computed. Let these lengths are a1, a2, a3 and so on as shown in fig 2. The thumb is excluded while analysing the hand using this approach.
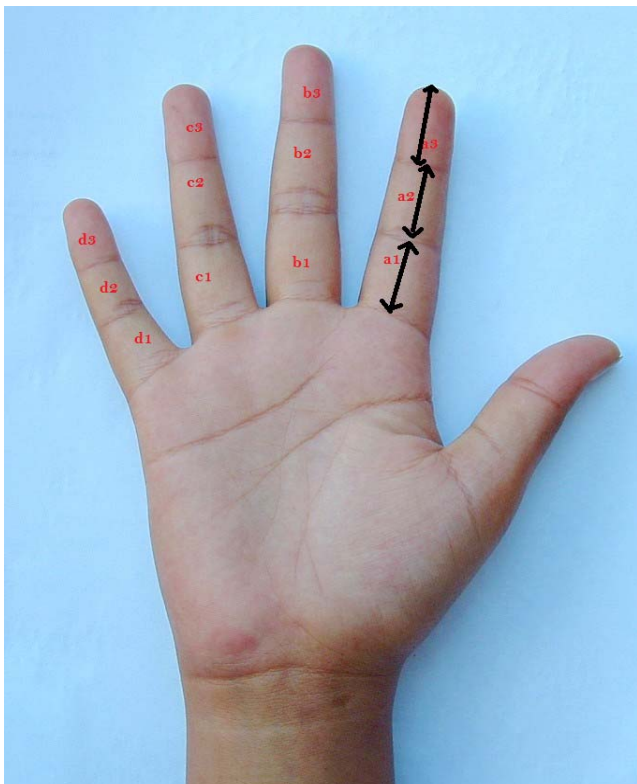
Fig. 2. Palm Image showing Finger's partition length

Using the partitions lengths following comparisons are made and thereof conclusions are drawn.

Case1: IF (a1+a2+a3) > (c1+c2+c3) && (a1+a2+a3) > (b1+b2+b3/2)
Conclusion :– Jupiter ruled person.

Case2: IF (c1+c2+c3) > (a1+a2+a3) && (c1+c2+c3) > (b1+b2+b3/2)
Conclusion: – Sun ruled person.

Case3: IF (a1+a2+a3) = (c1+c2+c3) && (a1+a2+a3), (c1+c2+c3) < (b1+b2+b3/2)
Conclusion: – Saturn ruled person.

Case4: IF (a1+a2+a3) = (c1+c2+c3) && (a1+a2+a3), (c1+c2+c3) > (b1+b2 + b3/2) && (d1+d2+d3) > (c1+c2 + c3/2)
Conclusion: – Mercury ruled person.

After analyzing the above results the following characteristics of different persons ruled by different planets are derived.

1. *Jupiter ruled-* leading ability, position seeking, want position first –money second

2. *Sun ruled-* creative, want a good public image, showy nature, and want to look good, flashy life style.

3. *Saturn ruled-* money minded, deep thinkers, love loneliness.

4. *Mercury ruled-* good speakers, tricky mind, business minded.

IV   EXPERIMENTAL RESULTS:

We implemented the proposed method in Matlab software. Palm image in jpeg format is taken as input (as shown in Fig 3). Edge detection is performed using canny filter to identify the edges of palm as shown in Fig 4. After the edge detection we matched the input image pattern with palm image templates from database to classify with which, the input image matches by maximum. Then we get a pattern of palm and palm boundaries to calculate desired lengths (Pw, Pl, Fl) by applying pixel by pixel matching with our database images. For images matching >70% is considered as suitable to generate desired pattern. After edge detection and pattern matching Hough Transform and pixel distance computation is applied, to compute palm-width (Pw), palm-length (Pl), and finger length (Fl) and their ratios (Pw/Pl and Fl/Pl). Once these parameters in form of Pw, Pl, Fl, Pw/Pl, and Fl/Pl are generated then the system takes the available information with respect to these data, already stored in database to generate personality type based on Palmistry facts. We tested our approach for palm images of well known personalities like Dalai Lama, Sir Arthur Salivan, Sara Bern Heart, William Whitley, General Sir Redvurse Buller, Benazir Bhutto and so on and found that the result produced by our system are

matching with the known characteristics about these personalities.

Fig. 3. Input Palm Image

Fig. 4. Output Palm Image showing Edges extracted using canny filter

V. CONCLUSION:

Palm length-width and finger length are very important characteristics of a palm and in this paper we have proposed a novel approach for getting palmistry results based on these characteristics with the help of image processing. In this paper we have proposed two approaches ratio based system approach and finger length comparison based approach. The ratio base approach helps to identify the basic characteristics of a person in the form of positive and negative characteristics whereas the finger length base approach help to identify different personality types ruled by different planets. With this approach one can make computer intelligent enough to read and analyze hands and generate results.

REFERENCE:

[1] Duda, Richard O., Hart, Peter E.: Use of Hough Transform to detect lines and curves. Stanford Research Institute. Communications of the ACM, Issue 1, (Nov1970).

[2] Richard O. Duda, Peter E. Hart: Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM, Volume 15, Issue 1, pp.11 – 15, January 1972.

[3] Xiangquan Wu, Kuanquan Wang, David Zhang: A Noel Approach of Palm Line Extraction. 3rd International Conference on image and graphics, 2004. Issue 1.

[4] Ugo Montanary, On The Optimal Detection Of Curves In The Noisy Pictures, Pisa, Italy. ACM Computer Survey. V.5, n.2, pp.81-88(1973)

[5] Stockman G.C., Agrawal A.K.: Equivalence of Hough Curve detection to template matching, L.N.K. Corporation, University Of Maryland. Communications of the ACM, v.20, n.11, pp.820-822(1977).

[6] History Of Palmistry, http://www.palmistry.com.au/history_of_palmistry.html

[7] Cairo.: Sampoorna Hast Rekha Vigyan. Manoj Publications (2004).

[8] William J. Benham.: Vrahad Hast Rekha Shastra. Manoj Publications (2007).

[9] Dayanand: Mysteries of Palmistry. Institute Of Palmistry(2009).

[10] Martin Van Mensvoort: Hands In News, http://www.handanalysis.com/

[11] 11.MaximumDistanceBetweenTwoPixels, http://www.mathworks.co.uk/matlabcentral/newsreader

AUTHORS PROFILE

Vishwa Ratna Nigam is undergraduate student in the discipline of Computer Science & Engineering of Jaypee Instittue of Information Technology, Noida.

Divakar Yadav is Assistant Professor in Jaypee Institute of Information Technology Noida, India. He received his undergraduate degree in Computer Science & Engineering form Univeristy of Lucknow, master degree in Information Technology from Indian Institute of Information Technology Allahabad, and Ph..D. degree in Computer Science & Engineering from Jaypee Institute of Information Technology, Noida, India. He has more than 15 research papers in reputed international/national journals and conferences.

Manish K Thakur is Senior Lecturer in Jaypee Institute of Information Technology Noida, India. He received his undergraduate degree in Elctronics Engineering form Nagpur Univeristy and master degree in Computer Science from Birla Institute of technology, Mesra Ranchi., India. He has more than 5 research papers in reputed international/national conferences.

Applying l-Diversity in anonymizing collaborative social network

G.K.Panda	A. Mitra	Ajay Prasad	Arjun Singh	Deepak Gour
Department of CSE & IT	Department of CSE & IT	Department of CSE	Department of CSE	Department of CSE
MITS, Sriram Vihar	MITS, Sriram Vihar	Sir Padampat	Sir Padampat	Sir Padampat
Rayagada, INDIA	Rayagada, INDIA	Singhania University,	Singhania University,	Singhania University,
gkpmail@sify.com	mitra.anirban@gmail.com	Udaipur, INDIA	Udaipur, INDIA	Udaipur, INDIA
		ajayprasadv@gmail.com	vitarjun@gmail.com	deepak.gour@spsu.ac.in

Abstract— **To date publish of a giant social network jointly from different parties is an easier collaborative approach. Agencies and researchers who collect such social network data often have a compelling interest in allowing others to analyze the data. In many cases the data describes relationships that are private and sharing the data in full can result in unacceptable disclosures. Thus, preserving privacy without revealing sensitive information in the social network is a serious concern. Recent developments for preserving privacy using anonymization techniques are focused on relational data only. Preserving privacy in social networks against neighborhood attacks is an initiation which uses the definition of privacy called k-anonymity. k-anonymous social network still may leak privacy under the cases of homogeneity and background knowledge attacks. To overcome, we find a place to use a new practical and efficient definition of privacy called l-diversity. In this paper, we take a step further on preserving privacy in collaborative social network data with algorithms and analyze the effect on the utility of the data for social network analysis.**

Keywords- bottom R-equal, top R-equal, R-equal, bottom R-equivalent, top R-equivalent and R-equivalent, l-diversity

I. INTRODUCTION

As the ability to collect and store more and more information about every single action in life has grown, huge amounts of details about individuals are now recorded in database systems. Social networks have always existed in society in varying forms. The record keeping power of computers and the advancement of internet, both the interactions within and scale of these social networks are becoming apparent. Individual social networks have proved fruitful and have been the topic of much research. However with the development of agencies, facilitators and researchers, the ability and desire to use multiple social networks collaboratively has emerged.

This has both positive and negative effects. The positive effects focus on many possibilities for enriching people's lives through new and improved social services and a greater knowledge of people's preferences and desires. The negative effect focuses on the concerns that private aspects of personal lives can be damaging if widely publicized. For example, knowledge of a person's locations, along with his preferences can enable a variety of useful location-based services, but public disclosure of his movements over time can have serious consequences for his privacy.

However, agencies and researchers who collect such data are often faced with a choice between two undesirable outcomes. They can publish personal data for all to analyze, this analysis may create severe privacy threats, or they can withhold data because of privacy concerns, this makes further analysis impossible and may hamper the social feel of the network and may lead to unpopularity of the site. Thus retaining individual privacy is really a concern to the social network analysis society.

A. Need of privacy in Social Network data

Let us ponder on two examples of social sharing that lead to troublesome situations.

Example 1: The Enron corporation bankruptcy in 2001 made available of 500,000 email messages public through the legal proceedings and analyzed by researchers [7]. This data set has greatly aided research on email correspondence, organizational structure, and social network analysis, but it also has likely resulted in substantial privacy violations for individuals involved.

Example 2: Network logs are one of the most fundamental resources to any computer networking security professionals and widely scrutinized in government and private industry. Researchers analyze internet topology, internet traffic and routing properties using network traces that can now be collected at line speeds at the gateways of institutions and by ISPs. These traces represent a social network where the entities are internet hosts and the existence of communication between hosts constitutes a relationship. Network traces (even with packet content removed) contain sensitive information because it is often possible to associate individuals with the hosts they use, and because traces contain information about web sites visited, and time stamps which indicate periods of activity.

There are five basic types of IP address anonymization algorithms in use. These are: black-marker anonymization, random permutations, truncation, pseudonymization and prefix-preserving pseudonymization. Somehow these are trivial and there are certain mapping methods which still have a chance to get exposed on attacks. To eliminate the hurdle of sharing logs, strong and efficient anonymization techniques are very much essential. [7].

Recent work has focused on managing the balance between privacy and utility in data publishing, but limited to relational

datasets. The definitions of k-anonymity [10] and its variants [2, 6] are promising. These techniques are commonly implemented with relational micro-data. While useful for census databases and some medical information, these techniques cannot address the fundamental challenge of managing social network datasets.

Bin Zhou and Jian Pei [15] proposed a privacy preservation scheme which deals against neighborhood attacks of social network using the definition of privacy called k-anonymity. However, k-anonymous social network still may leak privacy under the cases of homogeneity attacks and background knowledge attacks.

B. Contributions and Paper Outline

We propose an algorithm for the collaborative social network anonymization which can be extended to the higher level of security threat, where the adversary can have the information even about the vertices which are not the immediate neighbours of target vertex.

The basic definitions are provided in the next section. Section 3 describes the existing system to deal with the security in social network with k-anonymity as proposed by Zhou and Pei [15]. We extend the work to 2-neighborhood with an algorithmic approach and highlight possibilities of attacks. To overcome from such attacks in social network, we use a new practical and efficient definition of privacy called l-diversity [6]. It is proved [12] that l-diversity always guarantees stronger privacy preservation than k-anonymity. Section 4 highlights the proposed system of collaborative social network anonymization with equivalence relations and l-diversity method. Our goal is to enable the useful analysis of social network data while protecting the privacy of individuals. Finally section 5 gives the conclusion

II. DEFINITIONS AND NOTATIONS

In this section we reintroduce some basic notations that will be used in the remainder of the paper.

Definition-1 (Modelling Social Network) A social network can be modelled as a simple graph, $G = (V, E, L, \xi)$ where, V is the set of vertices of the graph, E is the edge set, L is the label set and ξ is the labelling function from vertex set V to label set L, $\xi = V \rightarrow L$.

Definition-2 (k-Anonymity) A table T satisfies k-anonymity if for every tuple $t \in$ T there exists k-1 other tuples ti1, ti2, ..., ti k-1 $\in$ T such that t[C] = ti1[C] = ti2[C] = ... = tik-1[C] for all C$\in$ QI

Theorem 1 (k-Anonymity): Let G be a social network and G' an anonymization of G. If G' is k-anonymous, then with the neighbourhood background knowledge, any vertex in G cannot be re-identified in G' with confidence larger than 1/k.

Definition-3 (Naive Anonymization) The naive anonymization of a graph $G = (V, E)$ is an isomorphic graph, Gna = (Vna, Ena), defined by a random bijection f: V-->Vna. The edges of Gna are Ena = {(f(x), f(x')) | (x, x')$\in$ E}

Definition-4 (Black-marker Anonymization) It replaces all IP addresses with a constant. It is quite similar to the affect as simply printing the log and blacking-out all IP addresses. This method is completely irreversible.

Definition-5 (Random permutation Anonymization) This method creates a one-to-one correspondence between anonymized and unanonymized addresses that can be reversed by one who knows the permutation.

Definition-6 (Truncation Anonymization) A fixed number of bits is decided upon (8, 16 or 24) and everything but those first bits are set to zero.

Definition-7 (Pseudonymization) It is a type of anonymization that uses an injective mapping such as random permutations.
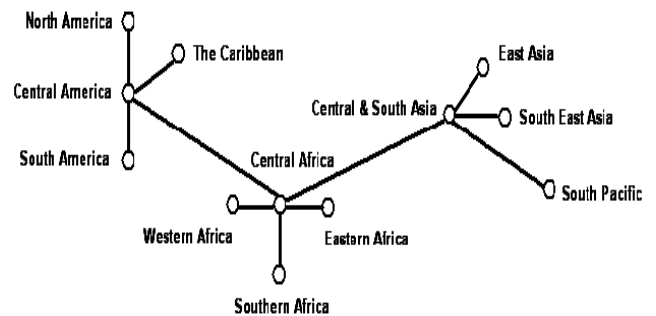
Figure 1. (a):A social network of interpol, G
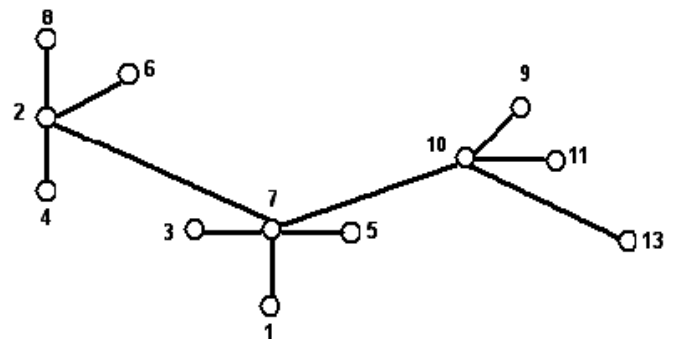
Figure 1. (b): The Naïve Anonymization of G

North America	8
Central America	2
South America	4
The Caribbean	6
Central Africa	7
Western Africa	3
Eastern Africa	5
Southern Africa	1
Central&South Asia	10
East Asia	9
South East Asia	11
South Pacific	13

Figure 1. (c): The Anonymization mapping

Definition-8 (Prefix-preserving Anonymization) In this anonymization, IP addresses are mapped to pseudo-random

anonymized IP addresses by a function where $\forall\, 1 \leq n \leq 32$, $Pn(x) = Pn(y)$ if and only if $Pn(\tau(x))=Pn(\tau(y))$.

III. THE EXISTING SYSTEM

The naive anonymization of a social network is to publish a version of the data that removes identification attributes. In order to preserve node identity in the graph of relationships, synthetic identifiers are used to replace them. *Figure 1* represents a social network of international police organization (Interpol), their naive anonymization and the anonymization mapping.

A. Publishing Social Network Data

To date publish of a giant social network jointly from different parties is an easier collaborative approach. Agencies and researchers who collect such social network data often have a compelling interest in allowing others to analyze the data. In many cases the data describes relationships that are private and sharing the data in full can result in unacceptable disclosures. Thus, preserving privacy without revealing sensitive information in the social network is a serious concern. Recent developments for preserving privacy using anonymization techniques are focused on relational data only

B. Preserving privacy in Social Networks using k-anonymity

The algorithm suggested by Bin and Jian [15] to anonymize a social network describes two basic steps as summarized below.

Step-1 Neighborhood Extraction and Vertex Organization

The neighborhood of each vertex is extracted and different components are separated. As the requirement is to anonymize all graphs in the same group to a single graph, isomorphism tests are conducted. For this purpose, for every component of the vertex the following steps are performed. Firstly all possible DFS trees are constructed for the component. Next, its DFS codes are obtained with respect to every DFS tree. Further the minimum DFS code is selected. This code is said to represent the component. Minimum DFS code has a nice property [14]: two graphs G and G0 are isomorphic if and only if DFS(G) = DFS(G0). Then neighborhood component code order is used to obtain single code for 1 vertex.

Step-2 Anonymization

Anonymization is done by taking the vertices from the same group. If the match is not found, the cost factor is used to decide the pair of vertices to be constructed.

Algorithm for k-Anonymization of one neighborhood

Input: A social network G=(V, E), the anonymization requirement parameter k, the cost function α, β and γ;

Output: An anomyzed graph G';

1: initialize G' = G;

2: mark $v_i \in V(G)$ as "unanonymized";

3: sort $v_i \in V(G)$ as VertexList in neighbourhood size - descending order;

4: WHILE (VertexList $\neq \phi$) DO

5: let SeedVertex = VertexList.head() and remove it from VertexList;

6: FOR each $v_i \in$ VertexList DO

7: calculate Cost(SeedVertex v_i) using the anonymization method for two vertices;

END FOR

8: IF (VertexList.size() $\geq$ 2k - 1) DO

let CandidateSet contain the top k - 1 vertices with the smallest Cost;

9: ELSE

10: let CandidateSet contain the remaining unanonymized vertices;

11: suppose CandidateSet= {u1,...um} anonymize

Neighbour(SeedVertex) and Neighbour(u1)

12: FOR j = 2 to m DO

13: anonymize Neighbour(uj) and {Neighbour(SeedVertex), Neighbour(u1) .... Neighbour(uj-1)}mark them as "anonymized";

14: update VertexList;

END FOR

END WHILE

C. Algorithm for k-Anonymization of two neighbourhoods

Step-1: Let u, v$\in$ V(G), u and v have similar neighbourhoods. Then, the labels are generalized or left unchanged, so that the neighbourhoods of u and v are isomorphic. Also, the labels of the vertices are same in both the neighbourhoods.

Step-2: If the neighbourhoods are not similar, the cost is calculated and the pair of vertices with the minimum cost is considered.

Step-3: The edges necessary are added to make them similar.

Step-4: The process of Step-1 is applied on the vertices pair.

D. Possible attacks on k- Anonymity
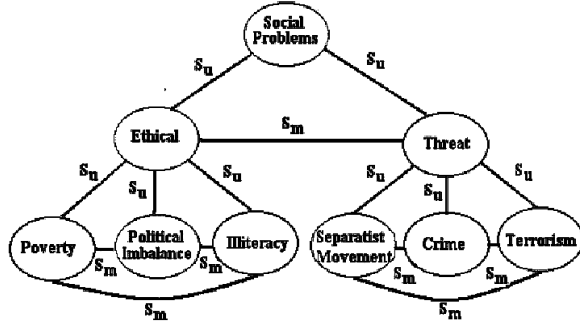
Figure 2. A social Network

Followings are two types of attacks that disclose sensitive information in k-anonymity [l-diversity] under two cases. First, if an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. Second, k-anonymity does not guarantee privacy against attackers using background knowledge (Attackers often have background knowledge).

Since both of these attacks are plausible in real life, it is required to go forward with the new definition of privacy that takes care of diversity and background knowledge.

Homogeneity Attack: k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute.

Background knowledge attack: k-Anonymity does not protect against attacks based on background knowledge.

IV. THE PROPOSED SYSTEM

In recent years, motivated by quality-aware scenarios like imprecise observations, there has been a growth of interest in models and algorithms for handling uncertain data, i.e. data describing many alternatives. Several working models of uncertain data have been proposed, which precisely describe many possible worlds by outlining the alternatives for possible events and the correlations or in dependencies between them. Given data presented in such models, there has been much effort in studying how to efficiently evaluate queries and perform analysis over the uncertain data and come up with a compact description of the possible answers to the queries.

It is quite clear to observe that, there is an important connection between the topics of Uncertain Data and Data Anonymization. The process of data anonymization introduces uncertainty into data that was initially certain. Data anonymization provides with principles methods for query evaluation and Uncertainty deals with a natural application area for uncertain data and both with a rich set of challenging problems.

A. Anonymity with structural equivalence

Nodes that look structurally similar may be indistinguishable to an adversary, in spite of external information.

Definition-3: In a social network, a pair of nodes x and y are said to be structurally equivalent [3] ($x \approx y$) when

1. $V (x, y) = \wedge (x , y)$

2. for any $v \in \wedge (x , y)$, $R (x, v) = R(y , v)$

3. if $y \in G (x)$, $R(x, y) = R(y , x)$

The equivalence ($x \approx y$) certainly means that x and y share a common set of relationships with a particular group of other nodes. It may not be necessary that x and y are to be directly connected/related.
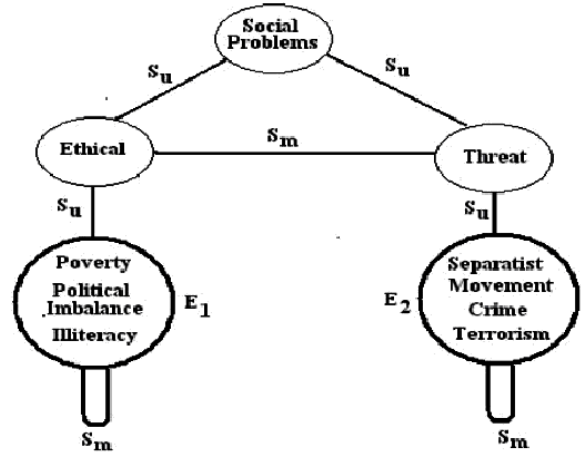
Figure 3. Reduction Social Network
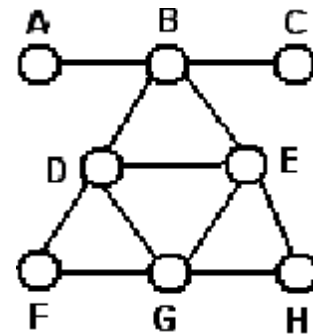
Figure 4. A Social Network Graph

Theorem 2: The equivalence relation induces a partition of N into disjoint equivalence classes satisfying:

1. x and y belong to the same equivalence class if and only if ($x \approx y$)

2. for $y \in G(x)$, R(x,y) is uniquely determined by the equivalence classes of x and y.

The key issue of this theorem is that the multitude of relations in the network can be summarized by simply

observing the relations that exist among equivalence classes. *Figure 2* shows a social network with a set of investigated social problems like Ethical Problems, Threats, Poverty, Political imbalance, Illiteracy, Separatist movement, Crime, Terrorism. *Figure 3* shows the reduction social networking with structural equivalence [3].

B. Automorphic Equivalence

Definition-9 (Automorphic equivalence): It is one of strong form of structural similarity between nodes. Two nodes x, y∈ V are automorphically equivalent (≈A) if there exists an isomorphism from the graph onto itself that maps x to y.

Automorphic equivalence induces a partitioning on V into sets whose members have identical structural properties. An adversary even with exhaustive knowledge of a target node's structural position, cannot isolate an individual beyond the set of entities to which it is automorphically equivalent. These nodes are structurally indistinguishable and observe that nodes in the graph achieve anonymity by being "hidden in the crowd" of its automorphic class members.

C. Vertex refinement

This technique originally developed to efficiently test for the existence of graph isomorphism. Here, the weakest knowledge query, H0 simply returns the label of the node. The queries are successively more descriptive, like H1(x) returns the degree of x, H2(x) returns the list of each neighbours' degree and so on. In general Hi(x) returns the multiple set of values which are the result of evaluating Hi-1 on the set of nodes adjacent to x: $Hi(x) = \{Hi-1(z1), Hi-1(z2),..., Hi-1(Zm)\}$ where z1,...zm are the nodes adjacent to x. Two nodes x, y in a graph are equivalent relative to Hi denoted x $\cong$ Hi, if and only if Hi(x) = Hi(y). Figure 4 is a simple social network graph, Table 1 represents the vertex refinement table and Table 2 is the equivalence class need to be further anonymized.

The proposed algorithm to anonymize a collaborated social network with l-diversity describes with three basic steps as summarized below.

Step-1 Formation of the collaborative social network by adding individual social networks

Step-2 Generating equivalence classes of nodes of the social network with vertex refinement having automorphic structural equivalence.

Step-3 Anonymization using l-diversity principle

D. The l-diversity Anonymization

Definition 10: (The l-diversity principle) A q*-block is l-diverse if contains at least l "well-represented" values for the sensitive attribute S. A table is l-diverse if every q*-block is l-diverse.

Definition 11: (The l-diversity principle in social network) An equivalence class of social network node implied by vertex refinement with structural equivalence is said to have l-

diversity if there are at least l "well-represented" values for the sensitive node. It is said to have l-diversity if every equivalence class has l-diversity.

E. Algorithm for Collaborative Social network with l-diversity

Input: A social network G = (V, E), l-diversity parameter

Output: An anonymized graph

1: S- The collaborative social network

2: S(o, e)

    o   – A node within the social network n

    o_j – The set of social networks that have provided with attributes, the numbering is made at the individual attribute level.

    e   - A set of edges related to node o

e_i – The set of social network from a user query

Sr   – The resulting social network from a user query

3: N – A social network is being added to S

   N(d, g) -

       d- A node within the social network

N

       g- A set of edges related to node o

4: R- A revocation social network being removed from S,

   R(d, g)

      d- A node within the social network R

      g- A set of edges related to node o

5: U- A user of S

   Uq – A query containing attributes to look for within S

   a(0)– An attribute or set of attributes of a node which can be used to uniquely identify the node

1: FOR each N(di, g)

2: IF (a(S(o)) = = a(N(d))) THEN

3: FOR each attribute within d

4: IF the attribute matches in o

5:   $o_j = o_j + N$

5: ELSE

6: ADD new attribute from d to o with $o_j = N$

7: END IF

8: FOR any edge within e

9: IF it matches an edge within g

10: $e_i = e_i + N$

11: ELSE

12: ADD non-matching edges within set g

    to set e with ei=N

    END IF

    END FOR

13: Add new node and edge set N(di, g) to S

    14: anonymize N(di) and N(g)

    END IF

    END FOR

V. CONCLUSIONS

In this paper we tried to focus the important problem of preserving privacy in publishing collaborative social network data which is really an important concern. We have referred the k-anonymity algorithm for social network using one neighbourhood and extended to two neighbourhood approach. Further we have discussed the relevant attacks to k-anonymity. We have studied the extent to which structural properties of a node using equivalence can serve as a basis of re-identification in anonymized social networks. This work further has potentiality to extend using the rich rough set theory. A l-diversity social network still may leak privacy. An adversary may have some prior belief about the sensitive attribute value of an individual before seeing the released table. After seeing the released table, the adversary may have a posterior belief. Information gain i,e., the difference between the posterior belief and the prior belief is the factor to leak privacy. Some mechanism analogous to t-closeness should be introduced.

REFERENCES

[1] Adam N. R. and Wortmann J.C., Security Control Methods for Statistical Databases: A Comparitive Study. ACM Computing Surveys, 21(4), pages 515-556, 1989.

[2] Bayardo, R.J. and Agarwal R., Data Privacy through Optimal k-Anonymization. In Proceedings of the IEEE International Conference of Data Eng., Washington, DC, USA. IEEE Computer Society, pages217-228 , 2005.

[3] Fiskel, J., Dynamic Evolution in Societal networks, J. Mathematical Sociology, pages 27-46, 1980.

[4] Johanne, S. and Pierre, M., Different relationships for coping with ambiguity and uncertainty in oganizations. J. Social Networks, Elsevier., pages 33-39, Vol 31, 2009.

[5] Li, N., Li, T. and Suresh, V., t-closeness: Privacy beyond k-anonymity and l-diversity. In ICDE, 2007.

[6] Machanavajjhala, A., Gehrek, J. and Kifer, D., l-diversity: Privacy beyond k-anonymity. In Proceedings of the 22nd Int. Conf. on Data Engg. ICDE 2006 (Atlanta, GA, USA, pages 24-29, April 03 - 08, 2006.

[7] McCallum. A., Corrade-Emmanuel, A., and Wang, X., Topic and role discovery in Social Networks, (2005). In IJCAI-05, 2005.

[8] Panda, G., K. and Mitra, A., Rough Set Application in Social Network using Rule Induction. In proceedings of the NCETIT, (2008), India, 2008.

[9] Panda, G. K. and Panda, B., S., Preserving Privacy in Social Networks with Covering Based Approximations of Classifications. In proceedings of NCACNIT-09. India, pages 525-530, 2009.

[10] Sweeny, L., K-anonymity: a model for protecting privacy. International Journal on uncertainty, Fuzziness and Knowledge based System, Pages 557-570, Vol 10(5), 2002.

[11] Wasserman, S. and Faust, K., Social Network Analysis. Cambridge University Press, 1994.

[12] Xiao, X. and Tao, Y., Personalized privacy preservation. In Proceedings of SIGMOD'06, 2006.

[13] Xu, J., Fan, J., Ammar, M. H. and Moon, S. B., On the Design and Performance of Prefix-Preserving IP Traffic Trace Anonymization. In ACM SIGCOMM Internet Measurement Workshop, 2001.

[14] Yan, X. and Han, J., gspan: Graph-based substructure pattern mining. In ICDM'02, 2002.

[15] Zhou, B. and Pei, J., Preserving Privacy in Social Networks Against Neighborhood Attacks. In Proceedings of the 24th Intl. Conf. On Data Engg. ICDE 2008, Cancun, Mexico, 2008.

3D-Mesh denoising using an improved vertex based anisotropic diffusion

Mohammed El Hassouni

DESTEC

FLSHR, University of Mohammed V-Agdal-
Rabat, Morocco

Mohamed.Elhassouni@gmail.com

Driss Aboutajdine

LRIT, UA CNRST

FSR, University of Mohammed V-Agdal-
Rabat, Morocco

aboutaj@fsr.ac.ma

Abstract—**This paper deals with an improvement of vertex based nonlinear diffusion for mesh denoising. This method directly filters the position of the vertices using Laplace, reduced centered Gaussian and Rayleigh probability density functions as diffusivities. The use of these PDFs improves the performance of a vertex-based diffusion method which are adapted to the underlying mesh structure. We also compare the proposed method to other mesh denoising methods such as Laplacian flow, mean, median, min and the adaptive MMSE filtering. To evaluate these methods of filtering, we use two error metrics. The first is based on the vertices and the second is based on the normals. Experimental results demonstrate the effectiveness of our proposed method in comparison with the existing methods.**

Keywords- **Mesh denoising, diffusion, vertex.**

I. INTRODUCTION

The current graphic data processing tools allow the design and the visualization of realistic and precise 3D models. These 3D models are digital representations of either the real world or an imaginary world. The techniques of acquisition or design of the 3D models (modellers, scanners, sensors) generally produce sets of very dense data containing both geometrical and appearance attributes. The geometrical attributes describe the shape and dimensions of the object and include the data relating to a unit of points on the surface of the modelled object. The attributes of appearance contain information which describes the appearance of the object such as colours and textures.

These 3D models can be applied in various fields such as the medical imaging, the video games, the cultural heritage... etc [1]. These 3D data are generally represented by polygonal meshes defined by a unit of vertex and faces. The most meshes used for the representation of objects in 3D space are the triangular surface meshes.

The presence of noise in surfaces of 3D objects is a problem that should not be ignored. The noise affecting these surfaces can be topological, therefore it would be created by algorithms used to extract the meshes starting from groups of vertices; or geometrical, and in this case it would be due to the errors of

measurements and sampling of the data in the various treatments [2].

To eliminate this noise, a first study was made by Taubin [3] by applying signal processing methods to surfaces of 3D objects. This study has encouraged many researchers to develop extensions of image processing methods in order to apply them to 3D objects. Among these methods, there are those based on Wiener filter [4], Laplacian flow [5] which adjusts simultaneously the place of each vertex of mesh on the geometrical center of its neighboring vertex, median filter [5], and Alpha-Trimming filter [6] which is similar to the nonlinear diffusion of the normals with an automatic choice of threshold.

The only difference is that instead of using the nonlinear average, it uses the linear average and the non iterative method based on robust statistics and local predictive factors of first order of the surface to preserve the geometric structure of the data [7].

There are other approaches for denoising 3D objects such as adaptive filtering MMSE [8]. This filter depends on the form [9] which can be considered in a special case as an average filter [5], a min filter [9], or a filter arranged between the two. Other approaches are based on bilateral filtering by identification of the characteristics [10], the non local average [11] and adaptive filtering by a transform in volumetric distance for the conservation of the characteristics [12].

Recently, a new method of diffusion based on the vertices [13] was proposed by Zhang and Ben Hamza. It consists in solving a nonlinear discrete partial differential equation by entirely preserving the geometrical structure of the data.

In this article, we propose an improvement of the vertex based diffusion proposed by Zhang and Ben Hamza. The only difference is to use of different diffusivities such as the functions of Laplace, reduced centred Gaussian and Rayleigh instead of the function of Cauchy. To estimate these various methods of filtering, two error metric L^2 [13] are used.

This article is organized as follows: Section 2 presents the problem formulation. In Section 3, we review some 3D mesh

denoising techniques; Section 4 presents the proposed approaches; Section 5 presents the used error metrics. In Section 6, we provide experimental results to demonstrate a much improved performance of the proposed methods in 3D mesh smoothing. Section 7 deals with some concluding remarks.

II. PROBLEM FORMULATION

3D objects are usually represented as polygonal or triangle meshes. A triangle mesh is a triple $M = (P, \varepsilon, T)$ where $P = \{P_1, ..., P_n\}$ is the set of vertices, $\varepsilon = \{e_{ij}\}$ is the set of edges and $T = \{T_1, ..., T_n\}$ is the set of triangles. Each edge connects a pair of vertices (P_i, P_j). The neighbouring of a vertex is the set $P^* = \{P_j \in P: P_i \sim P_j\}$. The degree d_i of a vertex P_i is the number of the neighbours P_j. $N(P_i)$ is the set of the neighbouring vertices of P_i. $N(T_i)$ is the set of the neighbouring triangles of T_i.

We denote by $A(T_i)$ and $n(T_i)$ the area and the unit normal of T_i, respectively. The normal n at a vertex P_i is obtained by averaging the normals of its neighbouring triangles and is given by

$$n_i = \frac{1}{d_{i5}} \sum_{T_j \in T(P_i^*)} n(T_j) \qquad (1)$$

The mean edge length $\bar{l}$ of the mesh is given by

$$\bar{l} = \frac{1}{|\varepsilon|} \sum_{e_{ij} \in \varepsilon} \|e_{ij}\| \qquad (2)$$

During acquisition of a 3D model, the measurements are perturbed by an additive noise:

$$P = P^{'} + \eta \qquad (3)$$

Where the vertex P includes the original vertex $P^{'}$ and the random noise process η. This noise is generally considered as a Gaussian additive noise.

For that, several methods of filtering of the meshes were proposed to filter and decrease the noise contaminating the 3D models.

III. RELATED WORK

In this section, we present the methods based on the normals such as the mean, the median, the min and the adaptive MMSE filters and the methods based on the vertices such as the laplacien flow and the vertex-based diffusion using the functions of Cauchy, Laplace, Gaussian and Rayleigh.

A. Normal-based methods

Consider an oriented triangle mesh. Let T and U_i be a mesh triangles, $n(T)$ and $n(U_i)$ be the unit normal of T and U_i respectively, $A(T)$ be the area of T, and $C(T)$ be the centroid of

T. Denote by $N(T)$ the set of all mesh triangles that have a common edge or vertex with T (see Fig. 1).
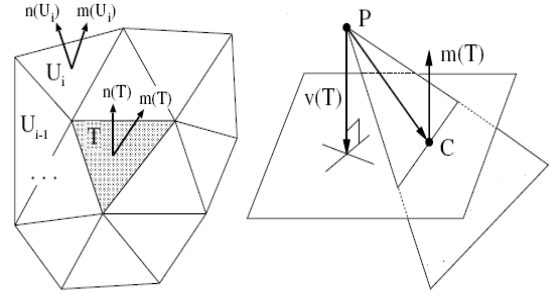
Fig. 1 Left: Triangular mesh.
Right: updating mesh vertex position.

1) Mean Filter: The mesh mean filtering scheme includes three steps [5]:

Step 1. For each mesh triangle T, compute the averaged normal $m(T)$:

$$m(T) = \frac{1}{\sum A(U_i)} \sum_{U_i \in N(T)} A(U_i) n(U_i) \qquad (4)$$

Step 2. Normalize the averaged normal $m(T)$:

$$m(T) \leftarrow \frac{m(T)}{\|m(T)\|} \qquad (5)$$

Step 3. Update each vertex in the mesh:

$$P_{new} \leftarrow P_{old} + \frac{1}{\sum A(T)} \sum A(T) v(T) \qquad (6)$$

With

$$v(T) = \left| \overrightarrow{PC}..m(T) \right| m(T) \qquad (7)$$

$v(T)$ is the projection of the vector $\overrightarrow{PC}..$ onto the direction of $m(T)$, as shown by the right image of Fig. 1.

2) Min filter : The process of min filtering differs from the average filtering only at step1. Instead of making the average of the normals, we determine the narrowest normal, η_i, for each face, by using the following steps [9]:

- Compute of angle Φ between $n(T)$ and $n(U_i)$.

- Research of the minimal angle: If Φ is the minimal angle in $N(T)$ then $n(T)$ is replaced by $n(U_i)$.

3) Angle Median Filter: This method is similar to min filtering; the only difference is that instead of seeking the narrowest normal we determine the median normal by applying the angle median filter [5]:

$$\theta_i = \angle\left(n(T), n(U_i)\right) \qquad (8)$$

If θ_i is the median angle in $N(T)$ then $n(T)$ is replaced by $n(U_i)$.

4) Adaptive MMSE Filter: This filter differs from the average filter only at step1. The new normal m(T) for each triangle T is calculated by [8]:

$$m_j(T) = \begin{cases} M_{lj}(T) & \sigma_n^2 > \sigma_{lj}^2 \, ou \, \sigma_{lj}^2 = 0 \\ \left(1 - \dfrac{\sigma_n^2}{\sigma_{lj}^2}\right)n_j(T) + \dfrac{\sigma_n^2}{\sigma_{lj}^2}M_{lj}(T) & \sigma_n^2 \le \sigma_{lj}^2 \, et \, \sigma_{lj}^2 \ne 0 \end{cases} \qquad (9)$$

$$M_{lj}(T) = \frac{\sum_{i=0}^{N-1} A(U_i) n_j(U_i)}{\sum_{i=0}^{N-1} A(U_i)} \qquad (10)$$

σ_n^2 is the variance of additive noise and σ_{lj}^2 is the variance of neighbouring mesh normals which is changed according to elements of normal vector. Thus, σ_{lj}^2 is calculated as follows:

$$\sigma_{lj}^2 = \frac{\sum_{i=0}^{N-1} A(U_i) n_j^2(U_i)}{\sum_{i=0}^{N-1} A(U_i)} - M_{lj}^2(T) \qquad (11)$$

B. Vertex-based methods

1) Laplacian Flow
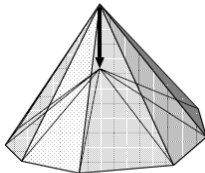
Fig 2. Updating vertex position by umbrella operator.

Considering the following expression which allows the update of the mesh vertices [5]

$$P_{new} \leftarrow P_{old} + \lambda D\left(P_{old}\right) \qquad (12)$$

Where $D(P)$ is a displacement vector, and λ is a step-size parameter.

The Laplacian smoothing flow is obtained if the displacement vector $D(P)$ is defined by the so-called umbrella operator [14] (see Fig. 2) :

$$U\left(P_i\right) = \frac{1}{n} \sum_{j \in N(P)} P_j - P_i \qquad (13)$$

$N(P)$ is the 1-ring of mesh vertices neighbouring on P_i.

2) Vertex-Based Diffusion using the Function of Cauchy:

This method [13] consists in updating the mesh vertices by solving a nonlinear discrete partial differential equation using the function of Cauchy.
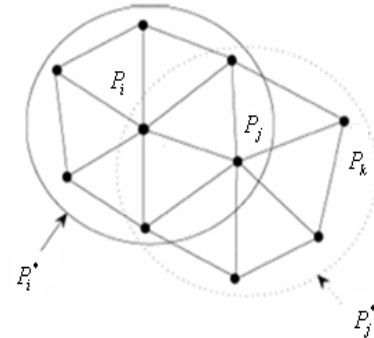
Fig 3. Illustration of two neighbouring rings.

The update of the vertices of mesh (see Fig. 3) is given by

$$P_i \leftarrow P_i + \sum_{P_j \in P_i^*} \frac{1}{\sqrt{d_i}} \left(\frac{P_j}{\sqrt{d_j}} - \frac{P_i}{\sqrt{d_i}} \right) \left(g\left(|\nabla P_i|\right) + g\left(|\nabla P_j|\right) \right) \qquad (14)$$

Where g is Cauchy weight function given by

$$g(x) = \frac{1}{1 + \dfrac{x^2}{c^2}} \qquad (15)$$

and c is a constant tuning parameter that needs to be estimated.

The gradient magnitudes are given by

$$|\nabla P_i| = \left(\sum_{P_j \in P_i^*} \left\| \frac{P_i}{\sqrt{d_i}} - \frac{P_j}{\sqrt{d_j}} \right\|^2 \right)^{1/2} \qquad (16)$$

And

$$|\nabla P_j| = \left(\sum_{P_k \in P_j^*} \left\| \frac{P_j}{\sqrt{d_j}} - \frac{P_k}{\sqrt{d_k}} \right\|^2 \right)^{1/2} \qquad (17)$$

Note that the update rule of the proposed method requires the use of two neighbouring rings as depicted in Fig. 3.

IV. PROPOSED METHOD

The method of vertex-based diffusion [13] was proposed by Zhang and Ben Hamza and which consists in solving a nonlinear discrete partial differential equation using the function of Cauchy.

In this section, we propose an improvement of the vertex-based diffusion proposed by Zhang and Ben Hamza. The only difference is the use of other diffusivity functions instead of Cauchy function. These functions are presented as follows:

- Reduced Centered Gaussian function :

$$g(x) = \sqrt{\frac{1}{2 \times pi}} \times \exp\frac{\left(-\left(\frac{x}{c}\right)^2\right)}{2} \qquad (18)$$

- Laplace function :

$$g(x) = \frac{\exp\left(-abs\left(\frac{x}{c}\right)\right)}{2} \qquad (19)$$

- Rayleigh function :

$$g(x) = \exp\left(-\frac{\left(\frac{x}{c}\right)^2}{2}\right) \times \left(\frac{x}{c}\right) \qquad (20)$$

c is a constant tuning parameter that needs to be estimated for each distribution.

V. L^2 ERROR METRIC

To quantify the better performance of the proposed approaches in comparison with the method based on the vertices using the function of Cauchy and the other methods, we computed the vertex-position and the face-normal error metrics L^2 [13].

Consider an original model M' and the model after adding noise or applying several iterations smoothing M. P is a vertex of M. Let set *dist (P, M')* equal to the distance between P and a triangle of the ideal mesh M' closest to P. Our L^2 vertex-position error metric is given by

$$\varepsilon_v = \frac{1}{3A(M)} \sum_{P \in M} A(P) dist(P, M')^2 \qquad (21)$$

Where $A(P)$ is the summation of areas of all triangles incident on P and $A(M)$ is the total area of M.

The face-normal error metric is defined by

$$\varepsilon_f = \frac{1}{A(M)} \sum_{T \in M} A(T) \|n(T') - n(T)\|^2 \qquad (22)$$

Here T and T' are triangles of the meshes M and M' respectively; $n(T)$ and $n(T')$ are the unit normals of T and T' respectively and $A(T)$ is the total area of T.

VI. EXPERIMENTAL RESULTS

This section presents simulation results where the normal based methods, the vertex-based methods and the proposed method are applied to noisy 3D models obtained by adding Gaussian noise as shown in Figs 6 and 8.
The standard deviation of Gaussian noise is given by

$$\sigma = noise \times \bar{l} \qquad (23)$$

Where $\bar{l}$ is the mean edge length of the mesh.

We also test the performance of the proposed methods on original noisy laser-scanned 3D models shown in Figs 4 and 10.
The method of vertex-based diffusion using the proposed diffusivity functions of Laplace, reduced centred Gaussian and Rayleigh are a little bit more accurate than the method of vertex-based diffusion using the function of Cauchy. Some features are better preserved with the approaches of vertex based diffusion using these functions (see Figs 4 and 10).
By comparing the four distinct methods (see Figs 5 and 11), we notice that the proposed method gives the smallest error metrics comparing to method of vertex-based diffusion using the function of Cauchy.
The experimental results show clearly that vertex-based methods outperform the normal-based methods in term of visual quality. These results are illustrated by Fig 6.
In Fig 7, the values of the two error metrics show clearly that the vertex-based diffusion using the functions of Laplace, reduced centred Gaussian and Rayleigh give the best results and they are more effective than the methods based on the normals. Fig 7 also shows that the approaches based on the

vertices such as Laplacien flow and the vertex-based diffusion using the functions of Cauchy, Laplace, reduced centred Gaussian and Rayleigh give results whose variation is remarkably small.

In all the experiments, we observe that the vertex-based diffusion using different laws is simple and easy to implement, and require only some iterations to remove the noise. The increase in the number of iteration involves a problem of over smoothing (see Fig 8). In Fig 9, we see that the method of vertex-based diffusion using the function of Cauchy leads more quickly to an over smoothing than the methods of vertex-based diffusion using the functions of Laplace, reduced centered Gaussian and Rayleigh.

VII. CONCLUSION

In this paper, we introduced a vertex-based anisotropic diffusion for 3D mesh denoising by solving a nonlinear discrete partial differential equation using the diffusivity functions of Laplace, reduced centered Gaussian and Rayleigh. These method is efficient for 3D mesh denoising strategy to fully preserve the geometric structure of the 3D mesh data. The experimental results clearly show a slight improvement of the performance of the proposed approaches using the functions of Laplace, reduced centered Gaussian and Rayleigh in comparison with the methods of the laplacien flow and the vertex-based diffusion using the function of Cauchy. The Experiments also demonstrate that our method is more efficient than the methods based on the normals to mesh smoothing.
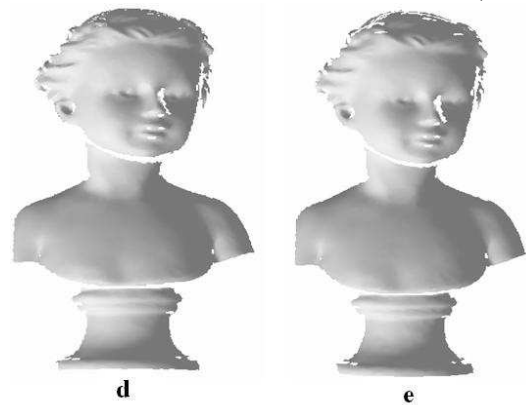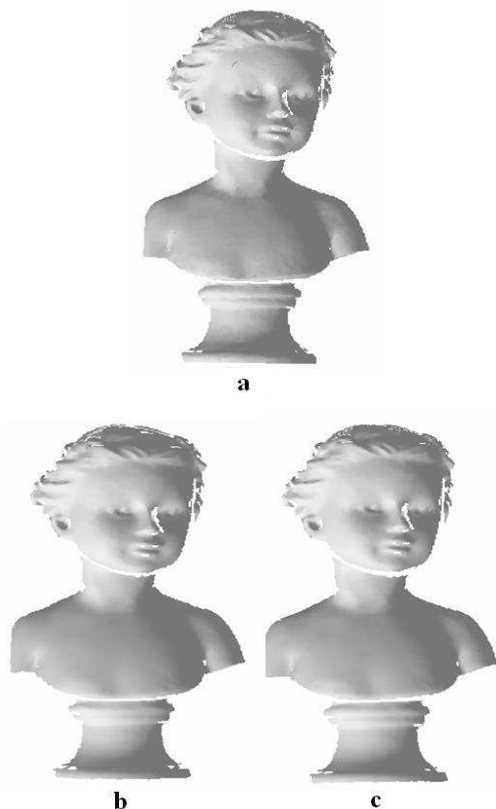
Fig 4. (a) Statue model digitized by a Roland LPX-250 laser range scanner (23344 vertices and 45113 faces); smoothing model by method based on the vertices using the functions of (b) Cauchy (c = 15.3849), (c) Laplace (c = 37.3849), (d) Gaussian (c = 37.3849) and (e) Rayleigh (c = 37.3849). The number of iteration times is 7 for each case.
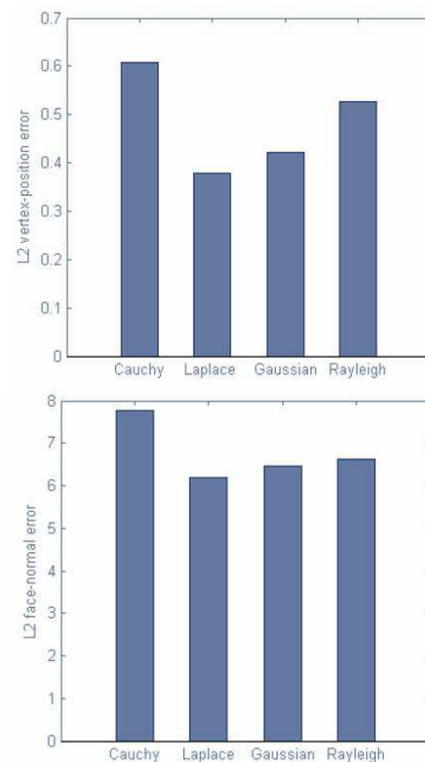
Fig 5. Top: L^2 vertex-position error metric of 3D model in Fig 4 Bottom: L^2 face-normal error metric of 3D model in Fig 4

Fig7. Left: L^2 vertex-position error metric of 3D model in Fig 6. Right: L^2 face-normal error metric of 3D model in Fig 6.

Fig 6. (a) Original model(4349 vertices and 2260 faces); (b) Adding Gaussian noise ($\varepsilon_v = 0.0090$, $\varepsilon_f = 0.0994$ and $\sigma = 0.8 \ \bar{l}$); (c) Min filter (7 iterations); (d) Mean filter (3 iterations); (e) Adaptatif MMSE filter (3 iterations); (f) Median filter (4 iterations); (g) Laplacien flow (2 iterations and λ=0.45); smoothing model by method based on the vertices using the functions of (h) Cauchy (3 iterations and c = 2.3849), (i) Laplace (6 iterations and c = 8.3849), (j) Gaussian (6 iterations and c= 8.3849) and (k) Rayleigh (6 iterations and c = 0.3).

Fig 8. (a) Original model (2108 vertices and 4216 faces); (b) Adding Gaussian noise ($\sigma = 0.7 \ \bar{l}$); smoothing model by method based on the vertices using the functions of (c) Cauchy (c = 2.3849), (d) Laplace (c = 15.3849), (e) Gaussian (c = 15.3849) and (f) Rayleigh (c = 0.03849). The number of iteration times is 10 for each case.
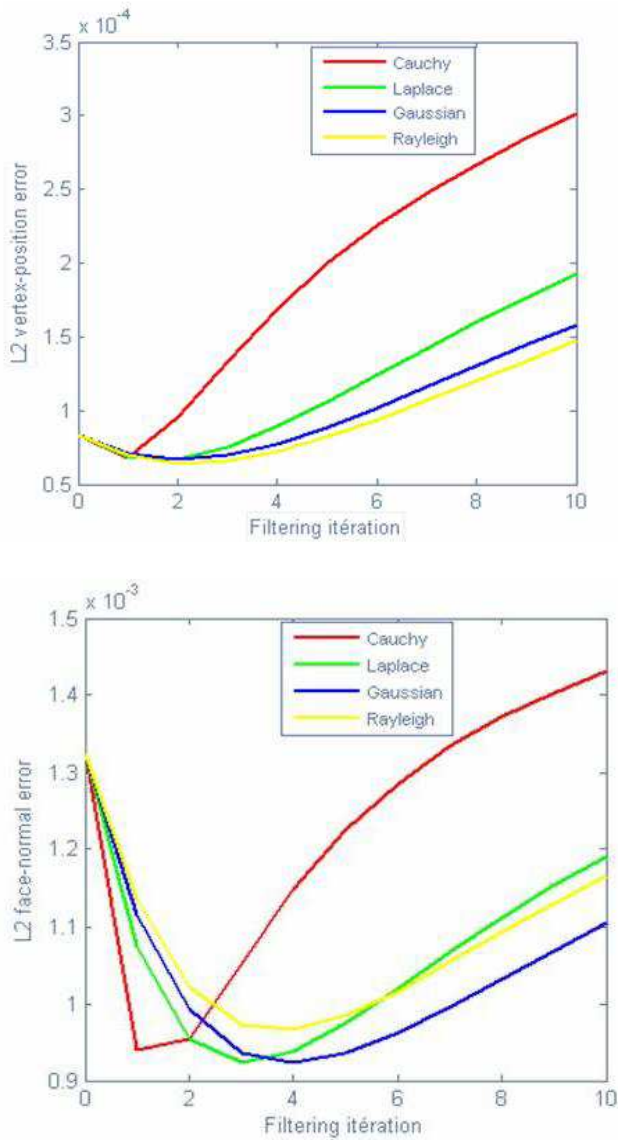
Fig 9. Left: L^2 vertex-position error metric of 3D model in Fig 8. Right: L^2 face-normal error metric of 3D model in Fig8.

Fig 10. (a) Statue model digitized by impact 3D scanner (59666 vertices and 109525 faces); smoothing model by method based on the vertices using the functions of (b) Cauchy (c = 15.3849), (c) Laplace (c = 37.3849), (d) reduced centered Gaussian (c = 37.3849) and (e) Rayleigh (c = 37.3849).The number of iteration times is 11 for each case.
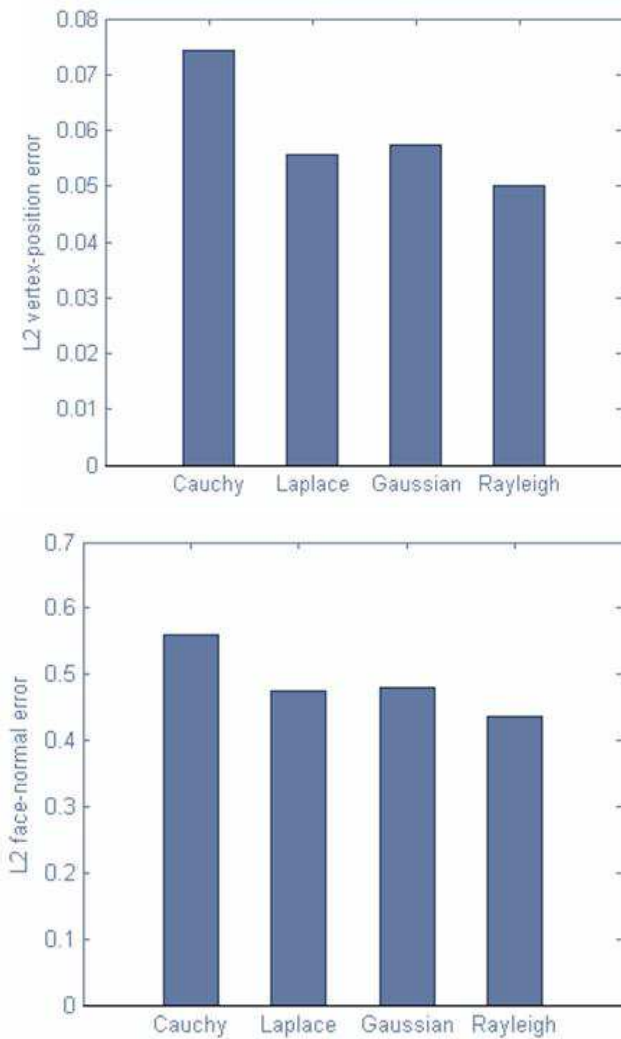
Fig 11. High: L^2 vertex-position error metric of Fig 10. Low: L^2 face-normal error metric of Fig 10.

REFERENCES

[1] Akram Elkefi et Marc Antonini, " Compression de maillages 3D multirsolution, transforme en ondelettes $2^{ème}$ génération ", rapport de recherche, 88 pages, novembre 2003.

[2] Michael Roy, "comparaison et analyse multirsolution de maillages irréguliers avec attributs d'apparence ", thèse de doctorat de l'Université de Bourgogne, 16 dcembre 2004.

[3] Gabriel Taubin, "A signal processing approach to fair surface design", International Conference on Computer Graphics and Interactive Techniques, Proceedings of the 22nd annual conference on Computer graphics and interactive techniques , SIGGRAPH, ACM Pages: 351 - 358, 1995.

[4] Jianbo Peng, Vasily Strela and Denis Zorin , "A Simple Algorithm for Surface Denoising", Proceedings of the conference on Visualization '01, pp. 107- 548, 21-26 October 2001.

[5] Hirokazu Yagou, Yutaka Ohtakey and Alexander Belyaevz, " Mesh Smoothing via Mean and Median Filtering Applied to Face Normals", Proceedings of the Geometric Modeling and Processing Theory and Applications, IEEE Computer Society, pp.124, 2002.

[6] Hirokazu Yagou, Yutaka Ohtake and Alexander G. Belyaev, "Mesh denoising via iterative alpha-trimming and nonlinear diffusion of normals with automatic thresholding", Proceedings of the Computer Graphics International (CGI'03) IEEE, pp. 28- 33, 9-11 July 2003.

[7] Thouis R. Jones, Frdo Durand and Mathieu Desbrun, "Non-iterative, feature-preserving mesh smoothing", Proceedings of ACM SIGGRAPH 2003, ACM Transactions on Graphics (TOG) Volume 22 , Issue 3 , pp. 943 - 949 , July 2003.

[8] Takashi Mashiko, Hirokazu Yagou, Daming Wei,Youdong Ding and Genfeng Wu, "3D Triangle Mesh Smoothing via Adaptive MMSE Filtering", Proceedings of the The Fourth International Conference on Computer and Information Technology (CIT'04) - Volume 00, pp.734 - 740 , 2004.

[9] Chen Chun-Yen and Cheng Kuo-Young, "A sharpness dependent filter for mesh smoothing", Computer Aided Geometric Design, Geometry processing, Volume 22, pp. 376-391, 2005.

[10] Takafumi Shimizu, Hiroaki Date, Satoshi Kanai, Takeshi Kishinami, "A New Bilateral Mesh Smoothing Method by Recognizing Features", Proceedings of the Ninth International Conference on Computer Aided Design and Computer Graphics (CAD-CG'05), IEEE Computer Society, pp 281-286 ,2005.

[11] Shin Yoshizawa, Alexander Belyaev and Hans-Peter Seidel, "Smoothing by Example: Mesh Denoising by Averaging with Similarity-based Weights", Proceedings of the IEEE International Conference on Shape Modeling and Applications (SMI'06), pp. 9, 14-16 June 2006.

[12] M. Fournier, J-M. Dischler et D. Bechmann, " Filtrage adaptatif des donnes acquises par un scanner 3D et représentées par une transformée en distance volumétrique ", Journées AFIG 2006, In Proceedings of FDB06a, pp. 171-178, Novembre 2006.

[13] Ying Zhang and A. Ben Hamza, "Vertex-Based Diffusion for 3-D Mesh Denoising", IEEE Transactions on image processing, Volume 16, n0 4, Avril 2007.

[14] L. Kobbelt, S. Campagna, J.Vorsatz, H.-P. Seidel, "Interactive multiresolution modeling on arbitrary meshes", ACM SIGGRAPH '98 proceedings, pp. 105-114, 1998.

[15] W. J. Rey, "Introduction to Robust and Quasi-Robust Statistical Methods". Berlin ; New York : Springer-Verlag, 19

A New Approach for Security Risk Assessment Caused by Vulnerabilities of System by Considering the Dependencies

Mohammad Taromi

Performance and Dependability Eng. Lab.
School of Computer Engineering, Iran University of
Science and Technology
Tehran, Iran
taromi@comp.iust.ac.ir

Mohammad Abdollahi Azgomi (Corresponding Author)

Performance and Dependability Eng. Lab.
School of Computer Engineering, Iran University of
Science and Technology
Tehran, Iran
azgomi@iust.ac.ir

Abstract— **Risk estimation is a necessary step in risk management which is the measurement of impact caused by the probability of exploiting vulnerabilities recognized in the system. At the moment, the qualitative metrics are used for this purpose that is believed to suffer subjectivity. The risk caused by a recognized vulnerability is computed using the values of common vulnerabilities scoring system (CVSS) attributes. But the great challenge in this field is that the dependency between vulnerabilities recognized in the system is not taken into account. In this paper, a new approach to risk assessment for the risks caused by vulnerabilities of system has been proposed which considers the dependencies among vulnerabilities. This approach consists of three steps. In the first step, after recognizing vulnerabilities of system and configuring the system, an attack graph is generated for all the critical resources of the system using MulVAL framework. Using these attack graphs, the dependency among vulnerabilities is extracted. In the second step, using the dependencies extracted among the vulnerabilities and estimated impact and exploitability defined based on CVSS attributes for individual vulnerability, a Markov model is generated. In the third step, using the Markov model, the quantitative security risk is estimated as the attacker keeps progressing in the system. In this paper we introduce the proposed approach, a case study demonstrating the above steps and the results of quantitative security risk estimation.**

Keywords-Security Risk Assessment; Vulnerability; Attack Graph

I. INTRODUCTION

Although engineering methods are applied in software production, with extending use and increasing complexities involved in information systems and market's requirements in reducing time and production costs, remarkable vulnerabilities remain unresolved in these systems. Furthermore, due to the intruders' different motivations in obtaining the resources of these systems or disturbing their functionality, the number of methods exploiting these vulnerabilities is also increasing. Despite the patching of vulnerabilities, due to the lack of appropriate patches, or the possibility of losing system's functionality after system reconfiguration, or even financial

limitations in providing patches for specific vulnerabilities, it is impossible to remove all these vulnerabilities. Moreover, despite using various attacker countermeasures such as firewalls or anti-viruses, the attackers are not easily recognized, or they are likely to disturb the system's ordinary operation. Therefore, due to the un-patched vulnerabilities and unrecognized attacks, there might be a security risk in system that should be managed [1, 28]. Thus, it is necessary for the administrator to manage the risk caused by these vulnerabilities. Risk estimation is a necessary step in risk management which is the measurement of impact caused by probability exploiting these vulnerabilities. Such estimation could be carried out either quantitatively or qualitatively. Estimating the quantitative risks using security metrics will be more useful than using qualitative metrics that are believed to suffer subjectivity [2].

Definition of vulnerability depends on the level of abstraction and the stage of system development. Vulnerability is an internal fault that empowers the external fault in damaging the system. In other words, vulnerability is of great importance in causing error and probably the resultant failure produced by the external fault [3]. The vulnerability addressed throughout this paper is based on the definition given by [4] as "a bug, flaw, weakness, or exposure of an application, system, device, or service that could lead to a failure of confidentiality, integrity, or availability". At the moment, it is possible to use open source scanners like OVAL [5] to recognize vulnerabilities in the host. The risk caused by a recognized vulnerability is computed using the values of common vulnerabilities scoring system (CVSS) attributes [4]. To do so, two components of risk assessment that are the exploitability and the impact due to the vulnerabilities are estimated. The advantage of using CVSS is that it employs a common open framework used by the experts for scoring and that it cannot be easily influenced by subjective judgment.

However, to evaluate the scoring of impact and exploitability in CVSS, the dependency between vulnerabilities recognized in the system is not taken into account [4]. To estimate the risk due to all vulnerabilities, it is necessary to take

the dependence between all vulnerabilities into consideration. By dependency, we mean that the possibility of exploiting vulnerability, after exploiting the other vulnerabilities, is taken into account. This dependency is usually modeled by attack graphs [6]. For this purpose, we have developed a dependency graph based on MulVAL [7] in which the exploitation of any vulnerability is possible by a certain privilege in the system. As a result of this exploitation, another privilege is provided for the attacker. The attacker attempts to obtain a critical privilege in the system. This graph is easy to understand in analyzing the vulnerabilities and has a lower presentational complexity then that generated in [8]. The study reported in this paper is an attempt to estimate the dependency between vulnerabilities in obtaining critical privilege by the attacker.

The impact of any vulnerability can be estimated based on the security properties (confidentially, availability, integrity), collateral damage potential (CDP) and distribution target (DT) by CVSS. A continuous-time Markov chain (CTMC) model is generated using the impact caused by the exploitability of any vulnerability by itself and the dependency obtained between the vulnerabilities in the system using the attack graph. In each state of this CTMC there are vulnerabilities whose impacts are similar. Categorizing these vulnerabilities in a particular state into groups is due to the fact that the attacker is charged by the minimum cost to obtain privilege or to manipulate the files or to deny services with similar impact. Moreover, the attacker does not try to exploit a series of vulnerabilities with similar impact. As a result, the dependency between these types of vulnerabilities in risk assessment is of little importance. In the proposed approach, the assumption is that there is not the possibility of repairing these vulnerabilities dynamically. As a result, it is not possible to transfer from one state with higher impact to another with lower impact. This assumption is completely logical. The reasons are as follows. First, risk assessment for a snapshot of the system is performed. Second, there is a meaningful time interval between the vulnerability recognized and offering a reliable path from software developer or it is not possible to patch the vulnerability because of interference. Having generated the model the quantitative risk assessment is estimated with attacker progress. Based on the results of this risk assessment, one can determine the best time to re-evaluate the system. It is worth to mention that model generation is become possible in a time complexity of $O(N^3)$, where N is the number of system states.

The advantages of the approach proposed in this paper are as follows. (1) It can be used to assess the risk caused by the threats from several critical parts of the system based on CVSS attributes for any vulnerabilities and dependency between them by considering the progresses of the attacker in the system. (2) In addition, it makes possible security evaluation of the system considering the data vulnerabilities and real environmental conditions to use the dependability techniques in security measurement. (3) It is possible to use the existing matured dependability evaluation techniques.

The rest of this paper is organized as follows. In section 2, the related works and their challenges and differences with this paper is discussed. In section 3, the existing methods of risk assessment for the risks resulting from any vulnerability are described using the values of CVSS attributes and new definitions for exploitability and impact of vulnerability are offered. Section 4, introduces how the dependency matrix is constructed based on the attack graph of the system. Section 5, presents how a Markov model is generated based on the dependency matrix, the impact and exploitability of vulnerabilities. Section 6, using the generated Markov model, the security risk of the system is estimated. Finally, in section 7, some concluding remarks are mentioned.

II. RELATED WORKS

In addition to quantitative and qualitative risk assessment, risk assessment methods are categorized into two groups: the first group (e.g. [9]), to which the method used in presented study belongs, takes into account all the possible sequences or the worst possible sequences as a basis for risk assessment considering all the vulnerabilities in the system and exploitability of them. The second group (e.g. [10]) operates taking into consideration the attacks succeeded which are gathered by intrusion detection system (IDS). The main advantage of the first category is that it takes into account all the possible sequences of exploitation. The second category, on the other hand, examines the attacker's behavior. However, due to false positive and false negative problems observed in alerts received from IDS, the state of system will not be precisely specified. Moreover, the more skilled intruders will display a different behavior because of their familiarity with how IDS operates. As a result, the estimated risk will have a lower reliability.

In [11] an initial model has been offered for quantitative measurement of security and the mean time and effort required for security breaches have been computed. This paper was one of first papers that put forward the idea of using dependability in security. The main challenge which using dependability analysis methods to achieve the security attributes of the system face is that in dependability analysis it is assumed that the failure occurred in the system or its components are random or rare events. However, in security analysis we are faced with failures caused by humans. The probability of such attacks depends on human beings' intelligent behavior and their learning through time [12].

In [13] the idea of using the attack graph to estimate the quantitative metric for the networks has been offered. This is akin to an often used metric of cryptographic strength which measures the weakest adversary who can break a cryptographic scheme. Since in attack graph to exploit a given vulnerability, certain conditions are required, these conditions cannot be achieved by attacker exploitation. Now, if the minimum required conditions to conduct exploitation in a network exceed those in a similar network but with a different configuration, it is clear that the first network can better fulfill security conditions than the second. In fact, this method has been offered to compare the similar networks with different configurations. Similar procedures are followed in [14] to hardening the network by achieving the minimum set of required conditions to close the paths with which the intruder tries to penetrate the system. In this paper, the severity of meeting all conditions were assumed to be the same. However,

the main challenge in such papers is that this problem is NP considering the very conditions.

The attack graph introduced in [15] whose nodes either describe the exploitation which are likely to be successful given all the conditions are met (as a result, it is called a AND node) or are pre- or post-conditions of the exploitations that could be assumed as OR nodes. According to the logic of these nodes, and using the intersection and conjunction operators corresponding to these nodes, and assuming that these conditions are independent of one another, and finally using the CVSS metric, the probability of reaching the target node examining all the paths available in the graph could be calculated. The difference between dependency graph generated in [16] and the one generated in the presented paper is that the graph offered in this study contains a vulnerability node that, if exploited, enables attacker's privilege. Therefore, the attack graph introduced in the current paper, the privilege, and vulnerability node follow the OR logic.

For the first time in [17], the idea of using web page ranking algorithm to score attack graph's nodes [18] was proposed. In this algorithm the significance of each node, like the webpage, depends on the number of paths the attacker could achieve. In [19], the changed web pages ranking algorithm has been applied onto the attack graph [8] that contains AND and OR nodes. In this way, the priority of each vulnerability for patching along with CVSS privilege is computed considering the dependency with other vulnerabilities. In our dependency graph the web page ranking algorithm can be employed but with fewer complexities.

In [9], the methodology for risk assessment of a potential threat which has been modeled using an attacker tree, first computes the dependency between the vulnerabilities to facilitate the exploitation of one vulnerability or another. Generating a dependency graph and the rate of facilitation between two vulnerabilities is determined by the expert. Using this dependency graph and the rate of facilitating each vulnerability based on such an updated exploitability and impact, the number of days when the service is not available has been defined, the risk resulting from each vulnerability has been estimated, and finally the total risk of threat has been estimated using the attack tree. The difference between the method used and one introduced in the present study lies in defining dependency. The dependency defined in [9] relies heavily on subjective judgment, whereas the dependency defined in the present paper is systematic that can be easily computed. In addition, in this paper, to estimate the exploitability and impact due to these vulnerabilities, the CVSS has been used. The approach taken is able to estimate the risk of several threats.

In [20], through combining the vulnerability attributes of CVSS using Bayesian networks, its impact and frequency have been estimated. Through combining these components, the resulting security risk has been computed. To achieve the total security for a given system, the use of Bayesian's algorithms has been suggested. In [21], a method has been offered to estimate the total security risk in a system. In this method, the vulnerabilities of the system have been divided into different

groups based on their impact. Then the groups have been ordered considering the impact of vulnerability. The system starts with a sound state until it encounters a failure. In our study, a different method of risk assessment has been proposed considering the dependencies between vulnerabilities.

TABLE I. CVSS METRIC GROUPS [4]

Base Metric Group		*Temporal Metric Group*	*Environmental Metric Group*
Access Vector(AV)	Confidentiality impact(B_C)	Exploitability(T_E)	Confidentiality(E_C), Integrity(E_I), Availability(E_A)
Access Complexity(AC)	Integrity impact(B_I)	Remediation Level (T_RL)	Collateral Damage Potential (CDP)
Authentication(Au)	Availability impact(B_A)	Report Confidence (T_RC)	Target Distribution (TD)

III. CALCULATING THE RISK OF ANY VULNERABILITY

CVSS [4] was introduced in 2004 and at the present second version is supported by Forum of Incident Response and Security Teams. It assigns a number to each vulnerability which is in vulnerability database like NVD [22]. In fact, CVSS is an open framework to determine the attribute and impact of vulnerability based on predefined and conceptable values to estimate the security risk due to this vulnerability. CVSS is consisted of three groups of metric: basic, temporary, and environmental.

The basic group metric is consisted of attributes that represent the inherent quality of vulnerability. The temporary group displays the attributes that changes over time and the environmental group shows those attributes that are unique to the user's immediate environment. The attributes of each group have been summarized in TABLE I. The metric for each group receives a value ranging from 0 to 10 and the content vector contains the values assigned to the attributes of the vulnerability that generate this numerical value.

CVSS offers a common set of attributes for vulnerabilities. All these attributes include presupposed qualitative values that are needed to select the values of the attributes of the vulnerability. For example, the attribute access vector from metric group that represents the way a vulnerability accessed and exploited, receives *L* value, this value indicates that the intruder is required to have physical access or a local account to exploit this vulnerability. The value of *A* suggests that the intruder should access local network of the host. Finally, the value of *N* indicates that the intruder can exploit the vulnerability without having a remote local access. To estimate the CVSS scores, for a given qualitative value a quantitative value has been assigned and using the equations that represent the relationships between these attributes, the basic group metric (the values of impact, and exploitability separately), the temporary group metric, and the environmental group metric(along with the adjusted impact) are estimated. To estimate these metrics, the CVSS calculator in NVD can be used. Due to the fact that in estimating basic exploitability in CVSS, the attributes of temporary group, all of which can affect the exploitability, are not considered, the exploitability addressed in this paper is defined as follows:

$$Exploitability = BaseExploitability(CVSS) * \\ T\_E*T\_RL*T\_RC \tag{1}$$

Furthermore, in the estimation of adjusted impact, in CVSS, two attributes of the environmental group, that is, collateral damage potential and target distribution are not taken into consideration. For this end, the impact addressed in this paper has been defined as follows (based on the metric of adjusted impact and the above-mentioned attributes):

$$Impact=2*AdjustedImpact(CVSS)*CDP*TD \tag{2}$$

The main problem, however, is that the total score or any generated metrics for each vulnerability by CVSS, or metric proposed by other methods [20], take into consideration the vulnerability by itself without reference to its dependency with other vulnerabilities.

IV. DEPENDENCY EXTRACTION AMONG VULNERABILITIES

A lot of studies have been conducted to generate the attack graph which shows all the sequences of exploitation of vulnerabilities in a network to attain critical privilege [6]. Recently, the challenge of many studies in this area has been to produce attack graphs with high scalability. But, to provide the data needed to better generate attack graphs, make it comprehensible, and its use for risk management networks are still hot topic in this field. In [8], using MulVAL [7], the logic based on the framework for vulnerability analysis, an algorithm to generate the attack graph with high scalability is presented. As a result, time complexity of attack graph generation has been reduced to quadratic time. The resultant attack graph has still presentational complexities which make it difficult to comprehend by humans. This challenge is discussed in [23] and the exploitation which do not provide deeper privilege on the network to attacker were removed.

MulVAL provides a framework based logic-programming approach to analyze multistage and multi-host attack path due to software vulnerabilities and misconfigurations. Network configurations, vulnerability specification, exploitation rules, and a set of privileges on network are specified by logic-programming language, Datalog. A logic program is a sequence of facts and rules. Facts are information about network elements, vulnerabilities, and privileges. Rules express how the attacker exploited existing facts to attain new facts about the network. Then, an off-the-shelf logic-programming engine that can evaluate logic-program efficiently in contrast with security policy violation which presents such "policy violation (Adversary, Access, Resource)", results in attack-traces of violation from security policy. Using this attack-traces attack graph is constructed.

Using OVAL Interpreter [5] vulnerability and the specific configuration are recognized. One major challenge in this field is the identification of pre and post-conditions of exploitation of vulnerabilities. Recently, in [24] on XML-based format similar to OVAL language has been proposed to express pre and post-conditions required for the exploitation of vulnerabilities with the purpose of using it in attack graphs

generation tools. In [25], approach for the extraction of this pre and post-condition of several vulnerability database is presented.
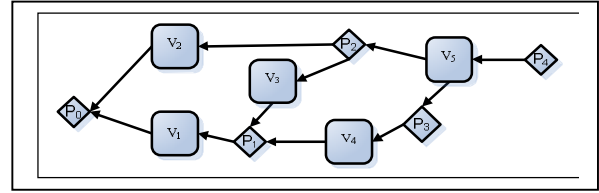
Figure 1.   Example of a proposed attack graph

Figure 2.   Dependency matrix of attack graph

To simplify the complexity of presentation of attack graphs, the attack graph which we generated using attack-traces output of MulVAL, only includes the vulnerabilities and privilege obtained from exploitation of these vulnerabilities. Each vulnerability can be exploited by one or more privileges. As a result, attacker will obtain a new privilege. Also, using one privilege, the attacker can exploit one or more vulnerabilities. For example, considering attack graph drawn in Figure 1, the attacker with privilege P_0 can exploit V_1 and V_2 vulnerabilities. His/her goal is to obtain the privilege P_4 on system. By dependency between V_1 and V_2 , we mean that exploiting V_2 provides a condition that enables exploiting V_1. For example, in Figure 1, the Vulnerabilities set $\{V_2, V_3, V_4\}$ which vulnerability V_5 depends on it, provides privileges P_2 and P_3 which enable the attacker in the exploitation of V_5. Dependency matrix $(|V| \times |V|)$ between vulnerabilities was extracted applying Breath First Search (BFS) on the generated attach graph. This dependency matrix belongs to the attack graph given in Figure 2.

V. THE PROPOSED METHOD

A system is often faced with vulnerabilities at any level of security. The intruder can decrease the level of service provided by the system through exploiting these vulnerabilities. The loss incurred as a result of service level drop which is imposed on the system in exploiting the vulnerabilities, depends on the collateral damage potential of the host where the vulnerabilities have been observed. In the initial state, the system includes all the vulnerabilities that can be exploited directly. The intruder decreases the level of the service provided by the system and it targets at a state where it can cause much security failure. Despite these attacks and drop of service level, the attacks can be tolerated by the system and the system manages to offer its main services accurately. As the exploitation proceeds, it provides the intruders with more opportunities to exploit the vulnerability causing sever security impact.   In addition, it makes the system enter a collateral

damage potential whose impact cannot be easily endured by the system or it is more likely to enter from these states into a complete failure state.

A. Generating the Markov Model

Having estimated the vulnerabilities recognized in the system, the impact and exploitability defined in section 2 for the vulnerability are estimated using CVSS attributes vulnerabilities. All the vulnerabilities except for the ones recognized in the initial state are categorized into N groups based on the impact they place on the system per se and the system's requirements. A group is a state of the system where the exploitation of any vulnerability recognized in the system has similar impact. The number of state (N) is equal to the number of mission tasks, or the number of user group's privilege, or the number of subsystems which can be attacked by the intruder. The justification for such a grouping is that the intruders select the easiest and most likely vulnerability to exploit the vulnerabilities that provide them with similar results. Moreover, because the vulnerabilities of a particular state result in a similar impact, the dependency between them is of little importance and they are not taken into account in risk assessment. In fact, this type of grouping is considered to be better than the grouping based on subsystems, privilege, and etc. to decrease the complexity involved in vulnerability analysis because it is conducted with reference to the component where the vulnerability is observed.

The transition rate between the two states is assumed to be the exploitability estimated of each vulnerability which is easy and more likely to exploit compared to the other vulnerabilities. Furthermore, the transfer between any two states occurs when the attacker can exploit the vulnerabilities of the new state. To achieve the transfer rate between these states, the dependency matrix introduced in section IV will be used. The assumption that intruder exploits the easiest vulnerability to transfer to another state provides the worst realistic estimation of the security risk and does not contradict the unpredictability of intruder's behavior.

The model generating algorithm has been generated using the impact due to the exploitation of vulnerabilities, the dependency matrix between them, and the threshold tolerable impact for the system. The vulnerabilities that are exposed to the attacker directly are categorized into the initial state, and the remaining vulnerabilities are categorized into the N states according to the impact due to them. This algorithm consists of the following four steps:

1) In the first step, for any state, the ascendents of each vulnerability in states which have fewer impact are extracted using the dependency matrix.

2) In the next step, the exploitability of vulnerabilities found in that state are normalized according to the exploitability of their ascendents, their own exploitability alone, and assuming that the exploitations of the ascendents of the vulnerabilities are independent of each other (for two ascendents shown in equation (3)). After estimating the exploitability of all vulnerabilities of a state, its exploitability is assumed to be equal to the probability of conjunction

exploitability of the vulnerabilities (*ExState*). As a result, a higher possibility of success for the attacker (a higher risk for the system) is taken into account in case there are more vulnerabilities in a state.

$$Ex(v) = (Ex(A) + Ex(B) + \frac{Ex(A)*Ex(B)}{10}) * \frac{Ex(v)}{10}; for \mid ascendents \, of \, v \mid = 2 \quad (3)$$

3) In the third step, the transfers between these states and their rates are determined such that. The attacker can transfer from S_i to S_j only if exploitability of at least one of the vulnerabilities of S_i allows the exploitation of at least one of the vulnerabilities in S_j.

4) And finally, from a state whose impact is above the tolerable threshold, a transfer is made to the failure state. In the process, to make sure that the states are reachable from the initial state, and it is possible to access the failure from any state, all the rows and columns of the transition-state matrix should be examined. In case there is not any transfer to any state except for the initial state, or there is no possibility of transfer to another state except for the failure state, the corresponding row and column of this state are removed from the transition-state matrix. Afterwards, the transition-state matrix is re-examined to ensure that such conditions are not present. In the worst case, the examination and removal of unreachable states from the initial state *N-1* is repeated. Moreover, the examination and removal of these states, has the complexity of $O(N^2)$.

The time complexity of an algorithm in proportion to the number of states is $O(N^3)$. The attacker will not transfer from a state with a higher impact to another with a lower impact. This is completely logical because the attacker is not naturally willing to transfer to a state where it has higher possible impact to another state where it has lower impact. In addition, as it was mentioned earlier, we do not consider dynamic reparability. Consequently, the resulting graph is directed acyclic graph (DAG).

VI. RISK ASSESSMENT

As it has been mentioned in many of the existing work on risk assessment, risk is the possibility of impact due to probability of exploiting the vulnerabilities in the system. It is obvious that the intruder should be able to access these vulnerabilities during risk assessment. Therefore, the risk due to the vulnerabilities that are not accessible to the intruder does not incur any risks to the system. For example, let us assume that the possibility of exploiting V_1 depends on exploiting V_2 and the mean time needed to exploit V_2 is t_1. Therefore, in t_1 time interval when the system keeps its initial conditions, V_1 poses no threat to the system. In addition, in t ($t > t_1$) risk does not involve exploiting V_2 because in the worst case, the failure due to the exploiting V_2 has been imposed on the system. As a result, the level of security provided by the system has decreased and there is more possibility that the intruders can exploit the vulnerabilities.

Generally, a limited number of vulnerabilities are exposed to exploitability in a second. As time passes, the intruder exploiting these vulnerabilities finds more opportunities to exploit vulnerabilities with sever impacts. This risk increases for a while. Finally, as the intruder exploits these vulnerabilities on critical hosts, the failure due to these vulnerabilities affects the system until it crashes. Thus, threat is posed by a risk, because there is no possibility of a further failure. Risk variations along time have been shown in Figure 3.
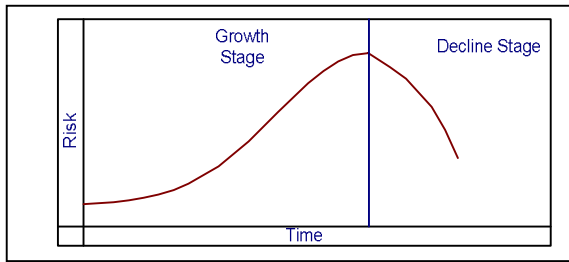
Figure 3. General diagram of system risk variation with time

For the estimation of system's risks, two components should be obtained: first, to consider the risk due to the vulnerabilities of a particular state in risk assessment, the mean time spent by the intruder to successfully exploit the vulnerabilities in this state should be calculated. Second, the states that are accessible by the intruder after a successful exploitation of a previous state should be obtained by examining the risk due to their vulnerabilities. To calculate the mean time elapsed at each state, according to [26], and since the Markov model utilized in this paper is an absorbing one where the states are divided into two groups of operational and faulty, the t vector is calculated as follows (the transition-rate matrix is limited to the operational states):

$$tQ = -p(0) \qquad (4)$$

Where Q is the transition-rate matrix restricted to operational states only, $p(0)$ is the initial state vector and p_i indicates the mean time takes the system to passes through the failure at the operational state i. After computing the mean time spent in each state to reach the failure state, using transition-rate matrix, it could be easily shown that when the intruder has reached the state i, which states are possible for the intruder to access as time t_i passes? In this way, the risk due to new states exposed to the intruder is taken into account. The total risk is estimated by the equation (5), regarding the change in reaching the states by the intruder. In this formula $impact_i$, the highest impact due to exploiting the vulnerabilities at the state i, and $ExState_i$, the exploitability of state i are included.

$$Risk(t) = \sum_{S_i \ accessible \ \& \ not \ exploited} impact_i * ExState_i \qquad (5)$$

VII. CASE STUDY

In this section, it will be shown how to estimate the risk applying the proposed approach on simple network given in [16]. In this network there are three hosts: Web server, File server, and Database server. For the attacker located in the internet only Web server is directly accessible. Firewall and network configuration determine reachability among hosts (in TABLE II). What vulnerabilities exist on any host, and the privilege required to exploit them (pre-condition), and the resultant privilege (post-condition) after the exploitation of vulnerabilities are given in TABLE III. All vulnerabilities on the network are remotely exploitable. For all vulnerabilities, CVSS attributes values of basic and temporal metric groups are gathered from NVD [22] and OSVDB [27]. Environmental security requirements in which the network is located are assumed to be similar to those of the network located in a university. Since, availability in this environment is very important, and integrity and confidentially are the next priorities, the cost of damage done to Database server is greater than those of servers and the cost of damage done to File server is greater than that of Web server. As a result, the value of CPD attribute of CVSS for vulnerabilities which is located in corresponding host is determined. Because the attacker can access the network via Web server value of TD attribute of CVSS for vulnerabilities located in Web server maximum possible value is selected. The Values of CVSS attributes for each vulnerability are given in TABLE III. In this table, exploitability and the impact of each vulnerability are also estimated.
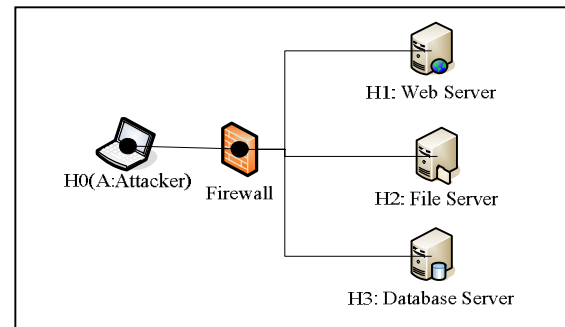
Figure 4. Configuration of example network

TABLE II. FIREWALL RULES OF NETWORK

Source	Dest.	Service	Action
All	H1	Http	Allow
All	H1	Ftp	Allow
All	H2	Ftp	Allow
H1	H3	Oracle	Allow
H2	H3	ftp	Allow

TABLE III.   VULNERABILITIES OF NETWORK AND ITS VALUES OF ATTRIBUTES OF CVSS

Vulnerability		Host Target	Pre & Post-condition		The Values of CVSS Attributes (B_Av/B_Ac/B_Au/B_C/B_I/B_A: T_E/T_RL/T_RC:E_C/E_I/E_A/CDP/TD)	Impact	Exploitability
			Pre	*Post*			
CVE-2002-0392	V₁	Web Server	Access≥user	Access=root	N/L/N/P/P:F/O/C:L/M/H/L/H	1.32	8.26
CVE-2003-1327	V₂	Web Server	Access≥user	Access=root	N/M/N/C/C/C:U/U/C:L/M/H/L/H	2.00	7.31
CVE-1999-0017	V₃	Web Server	Access=user	Access=user	N/L/N/P/P:H/N/N:L/M/H/L/H	1.32	10.00
CVE-1999-0017	V₄	File Server	Access=user	Access=user	N/L/N/P/P:H/N/N:L/M/H/Mh/M	3.96	10.00
CVE-1999-0017	V₅	DataBase Server	Access=user	Access=user	N/L/N/P/P:H/N/N:L/M/H/H/M	4.95	10.00
CVE-2001-0499	V₆	DataBase Server	Access≥user	Access=root	N/L/N/P/P:H/N/N:L/M/H/H/M	7.50	7.39

A. **Step1**: Calculating Dependency Among Vulnerabilities

As mentioned before, dependency among vulnerabilities are extracted from attack graph. To generate attack graph with properties noted in section IV, first reachability and vulnerability specification as input of MulVAL are extracted. The attack graph in Figure 7 is generated for "execCode(dbServer , root)" violation from security policy. In generation of this attack graph all none simple path, mentioned in [23], for easier presentation. Since this network is very simple, it is clearly understandable that attack graph generation for other goals do not add extra dependency to the dependency matrix. As a result, the dependency matrix of network is presented in Figure 5.

$$
\begin{array}{c}
\begin{array}{cccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \end{array} \\
\begin{array}{c} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{array}
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0
\end{bmatrix}
\end{array}
$$

Figure 5.   Dependency Matrix is extracted from attack graph of Figure 7

B. **Step2**: Create Purposed Model

In initial state were categorized V_1 and V_2 vulnerabilities because directly reachable for attacker. Difference among impact value of vulnerabilities considered one, as a result three states are obtained for the model. The complete failure state occurs when the root privilege of Database server is obtained by attacker. The Markov model created by algorithm for the network is presented in Figure 6. On the graph model transition rate among states is given.
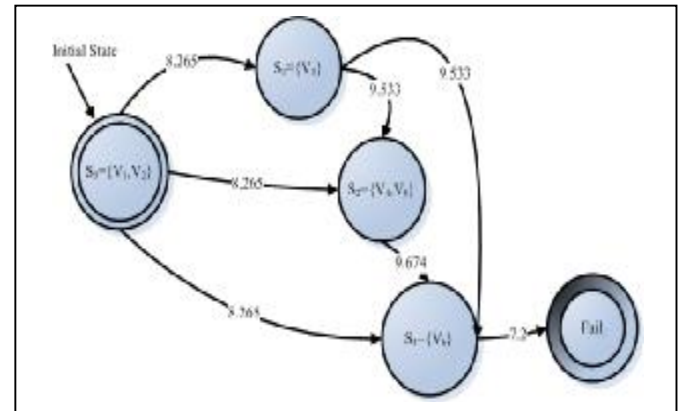
Figure 6.   The markov model generated for the network

TABLE IV.   RESIDENCE TIME FOR EACH STATES

State	Impact	τᵢ	Exploitability of State
S_0	2.00	0.040	9.533
S_1	1.32	0.016	9.533
S_2	4.95	0.050	9.989
S_3	7.50	0.136	7.200

C. **Step3**: Risk Assessment

With restrict transition-rate matrix to states except the failure state and considering the state S_0 as initial state using equation (4) can compute mean resident time in each state until achieve failure state. The values of impact, exploitability and mean resident time ($τ_i$) in each state is presented. Using the values of TABLE IV and the risk defined in equation (5) is estimated with attacker progress in network and is presented in Figure 8. Considering the Figure 8 one can understand that with successful exploitation of V_1 and V_2 the vulnerabilities by attacker, the risk extremely increases.
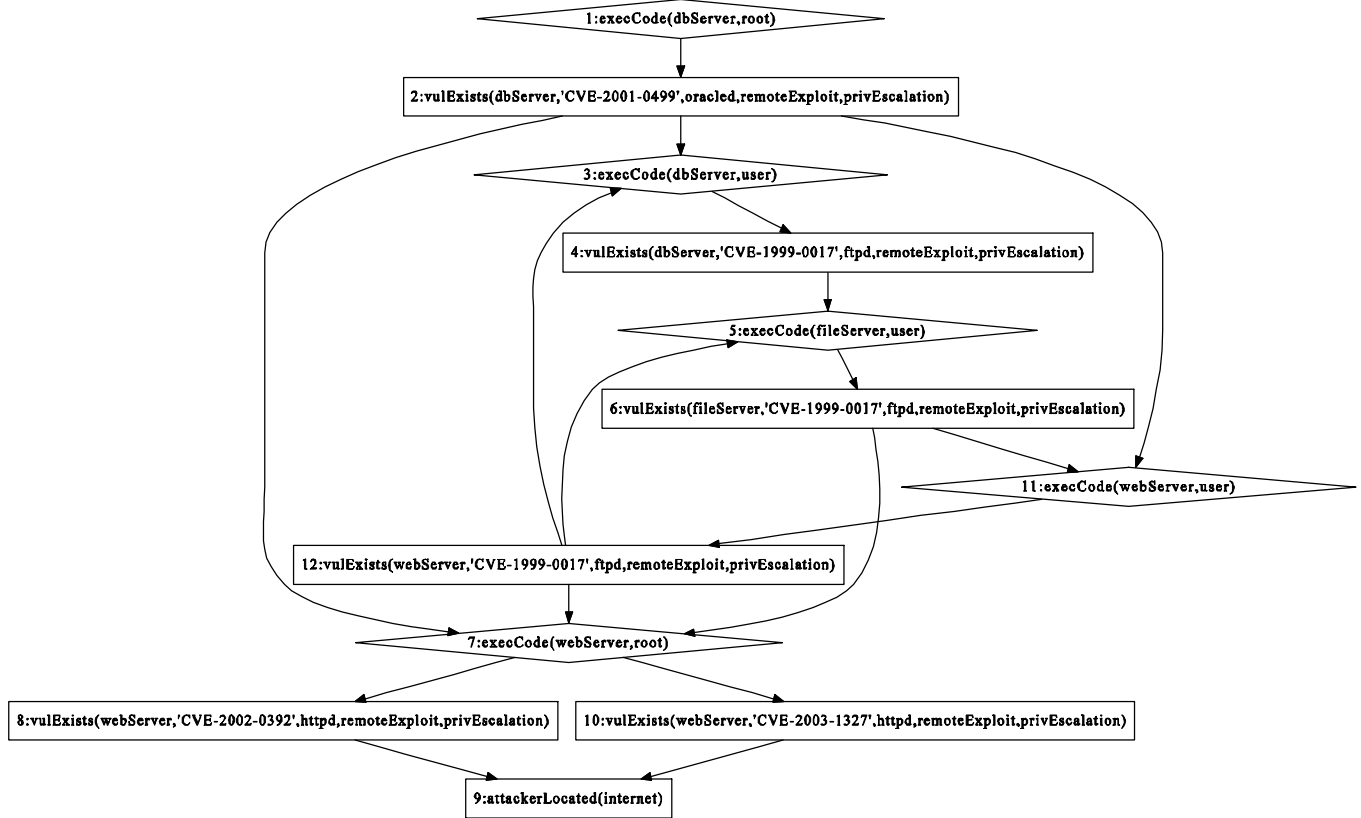
Figure 7. The attack graph of the example network for DB Server
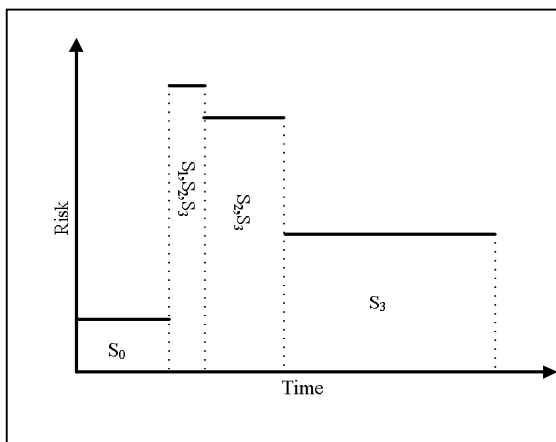
Figure 8. Risk variation with attacker progress in the network

VIII. CONCLUSIONS AND FUTURE WORKS

In this paper, we presented a new approach for estimating the overall security risk due to system vulnerabilities with regard to the dependency between them. First, the dependencies between the vulnerabilities are extracted from attack graphs of the system. Based on these dependencies and the impact of exploitation of vulnerabilities a Markov model was presented. In addition, this model provides the possibility of using mature dependability techniques for security measurement that can be used for simplifying attack graph presentation with removing "Useless" exploitations which do not provide deeper access in network for the attacker. Since the complexity of the model generation offered is obtained in $O(N^3)$, this method is very better than the method presented in MulVAL with exponential time. In the method proposed in this paper all directly reachable vulnerabilities were categorized into initial state because of the limitation of simple Markov. In future works we will extend the model and consider the attackers that can start from each state with some probability.

REFERENCES

[1] G. Stoneburner, A. Goguen, and A. Feringa,"Risk management guide for information technology systems," National Institute of Standards and Technology, special publication 800-30, 2002.

[2] M. Sahinoglu,"An input–output measurable design for the security meter model to quantify and manage software security risk," IEEE Transactions On Instrumentation And Measurement, June 2008, Volume: 57, pp. 1251-1260.

[3] A. Avizienis, J. C. Laprie, B. Randell, and C. Landwehr,"Basic concepts and taxonomy of dependable and secure computing," IEEE Transactions On Dependable And Secure Computing, Vol. 1, No. 1, January-March 2004, pp. 11-33.

[4] common Vulnerability Scoring System(CVSS), http://www.first.org/cvss/cvss-guide.html (3/1/2010)

[5] Open Vulnerability and Assessment Language (OVAL). http://oval.mitre.org/index.html (3/1/2009)

[6] R. Lippmann, and et al, "An annotated review of past papers on attack graphs - pr-ia-1", *MIT Lincoln Laboratory Project Report*, 31 March 2005.

[7] X. Ou, A logic-programming approach to network security analysis. Ph.D. thesis, Princeton University, 2005.

[8] X. Ou, W. F. Boyer, and M. A. McQueen, "A scalable approach to attack graph generation," *In Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS 2006)*, Alexandria, VA, U.S.A., October 2006. pp. 336 – 345.

[9] M. Benini, S. Sicari, "Risk assessment in practice: a real case study," *Computer Communications*, Vol. 31 (issue 15), September 2008, pp. 3691-3699.

[10] Arnes, K. Sallhammar, K. Haslum, T. Brekne, M. Moe, and S. J. Knapskog, "Real-time risk assessment with network sensors and intrusion detection systems," *In Proceedings of the International Conference on Computational Intelligence and Security* (CIS'05), Xian, China, December 2005, LNCS Vol. 3802, pp. 388-397

[11] Littlewood, S. Brocklehurst, N. Fenton, P. Mellor, S. Page, and D. Wright, "Towards operational measures of computer security," Journal of Computer Security, 1993,Vol. 2, pp. 211-229.

[12] M. Nicol, W. H. Sanders, and K. S. Trivedi, "Model-based evaluation: from dependability tosecurity," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, issue 1, Jan 2004, pp. 48-65.

[13] J. Pamula, S. Jajodia, P. Ammann, and V. Swarup, "Network security metrics: A weakest adversary security metric for network configuation security analysis," *In Proceedings of the 2nd ACM Workshop on Quality of Protectio,* Alexandria, VA, USA, October 2006, pp.31-38.

[14] L. Wang, N. Steven, and S. Jajodia, "Minimum-cost network hardening using attack graphs," *Computer Communications*, Vol. 29, Issue 18 , 28 November 2006, pp. 3812-3824.

[15] L. Wang, A. Singhal, and S. Jajodia, "An attack graph-based probabilistic security metric," *In Proceedings of the 22nd Annual IFIP WG 11.3 Working Conference on Data and Applications Security (DBSEC 2008)*, London, U.K., July 13-16, 2008, LNCS, Vol. 5094, pp. 283-296.

[16] P. Ammann, D. Wijesekera, and S. Kaushik, "Scalable, graph-based network vulnerability analysis," *In Proceedings of The 9th ACM Conference on Computer and Communications Security,* Washington, DC, November 2002, pp. 217-224.

[17] V. Mehta, C. Bartzis, H. Zhu, E. Clarke, and J. Wing, "Ranking attack graphs," *In The Proceedings of Recent Advances in Intrusion Detection (RAID),* Massachusetts, USA, September 2006, LNCS, Vol. 4219, pp. 127-144.

[18] O. Sheyner, J. Haines, S. Jha, and R. Lippmann, and J. M. Wing, , "Automated generation and analysis of attack graphs," *In Proceedings of the IEEE Symposium on Security and Privacy*, Oakland, CA, May 2002, pp. 273-284.

[19] R. Sawilla and X. Ou, "Identifying critical attack assets in dependency attack graph," *In Proceedings of the 13th European Symposium on Research in Computer Security (ESORICS)*,Malaga, Spain,October 2008, LNCS, Vol. 5283, pp. 18-34.

[20] S. H. Houmb, V. Nunes Leal Franqueira, and E. A. Engum, ,"Quantifying security risk level from cvss estimates of frequency and impact," *Journal of systems and software*, ISSN 0164-1212. (in press)

[21] S. H. Houmb, and V. Nunes Leal Franqueira, "Estimating ToE risk level using CVSS," *In Proceedings of the Fourth International Conference on Availability, Reliability and Security* (ARES 2009 The International Dependability Conference), 16-19 March 2009, Fukuoka, Japan, IEEE Computer Society, pp. 718-725.

[22] NIST, National Vulnerability Database, NVD, http://nvd.nist.gov/ (accessed March 3, 2010)

[23] J. Homer, A. Varikuti, X. Ou, and M. A. McQueen, "Improving attack graphvisualization through data reduction and attack grouping," *Proceedings of the 5th International Workshop on Visualization for Cyber Security (VizSEC)*, Cambridge, MA USA, September 15, 2008, LNSC, Vol. 5210, pp. 6879.

[24] P. Maggi, P, D. Pozza, and D. Sisto, "Vulnerability modelling for the analysis of network attacks," *In Proceedings of the Third International Conference on Dependability of Computer Systems DepCoS-RELCOMEX*, Washington, DC, USA, ,2008, IEEE Computer Society, pp.15-22.

[25] S. Roschke, F. Cheng, R. Schuppenies, and C. Meinel, "Towards unifying vulnerability information for attack graph construction," *In Proceedings of the 12th International Conference on Information Security,* Pisa, Italy, 2009, LNCS, Vol. 5735, pp. 218.233.

[26] K. S. Trivedi, Probability and Statistics with Reliability, Queuing, and Computer Science Applications, John Wiley and Sons, New York, 2001. ISBN number 0-471-33341-7.

[27] OSVDB: The Open Source Vulnerability Database, osvdb.org, *(accessed March 3, 2010)*

[28] K. Scarfone, and T. Grance, "A framework for measuring the vulnerability of hosts," *In Proceedings of the 1st International Conference on Information Technology*, IEEE Computer Society, pp. 1-4**.**

AUTHORS PROFILE

Mohammad Taromi is currently M.Sc. student in computer engineering (software) at school of computer engineering, Iran University of Science and Technology, Tehran, Iran.

His research interests include network security, vulnerability analysis, security estimation and evalution, and modelling and analysis of dependable system.

Mohammad Abdollahi Azgomi received the B.S., M.S. and Ph.D. degrees in computer engineering (software) (1991, 1996 and 2005, respectively) from Sharif University of Technology, Tehran, Iran.

His research interests include performance and dependability modelling with high-level modelling formalisms such as stochastic Petri nets, tools for modelling and evaluation, verification and validation, object-oriented modelling, web services, grid computing and network security. He has published several papers in international journals and conferences.

Dr. Abdollahi Azgomi is currently a faculty member at the school of computer engineering, Iran University of Science and Technology, Tehran, Iran.

IMAGE SUPER RESOLUTION USING MARGINAL DITRIBUTION PRIOR

S.Ravishankar

Department of Electronics and Communication
Amrita Vishwa Vidyapeetham University
Bangalore, India
s_ravishankar@blr.amrita.edu

Dr.K.V.V.Murthy

Department of Electronics and Communication
Amrita Vishwa Vidyapeetham University
Bangalore, India
kvv_murthy@blr.amrita.edu

Abstract— **In this paper, we propose a new technique for image super-resolution. Given a single low resolution (LR) observation and a database consisting of low resolution images and their high resolution versions, we obtain super-resolution for the LR observation using regularization framework. First we obtain a close approximation of the super-resolved image using learning based technique. We learn high frequency details of the observation using Discrete Cosine Transform (DCT). The LR observation is represented using a linear model. We model the texture of the HR image using marginal distribution and use the same as priori information to preserve the texture. We extract the features of the texture in the image by computing histograms of the filtered images obtained by applying filters in a filter bank and match them to that of the close approximation. We arrive at the cost function consisting of a data fitting term and a prior term and optimize it using Particle Swarm Optimization (PSO). We show the efficacy of the proposed method by comparing the results with interpolation methods and existing super-resolution techniques. The advantage of the proposed method is that it quickly converges to final solution and does not require number low resolution observations.**

Keywords-component; formatting; style; styling; insert (key words)

I. INTRODUCTION

In many applications high resolution images lead to better classification, analysis and interpretation. The resolution of an image depends on the density of sensing elements in the camera. High end camera with large memory storage capability can be used to capture the high resolution images. In some applications such as wildlife sensor network, video surveillance, it may not be feasible to employ costly camera.
In such applications algorithmic approaches can be helpful to obtain high resolution images from low resolution images obtained using low cost cameras. The super-resolution idea was first proposed by Tsai and Huang [1]. They use frequency domain approach and employ motion as a cue. In [2], the authors use a Maximum a posteriori (MAP) framework for jointly estimating the registration parameters and the high-resolution image for severely aliased observations. The authors in [3] describe an MAPMRF based super-resolution technique using blur cue and recover both the high-resolution scene intensity and the depth fields simultaneously. The authors in [4] present technique of image interpolation using

wavelet transform. They estimate the wavelet coefficients at higher scale from a single low resolution observation and achieve interpolation by taking in-verse wavelet transform. The authors in [5] propose technique for super-resolving a single frame image using a database of high resolution images. They learn the high frequency details from a database of high resolution images and obtain initial estimate of the image to be super-resolved. They formulate regularization using wavelet prior and MRF model prior and employ simulated annealing for optimization. Recently, learning based techniques are employed for super-resolution. Missing information of the high resolution image is learned from a database consisting of high resolution images. Freeman et al. [6] propose an example based super-resolution technique. They estimate missing high-frequency details by interpolating the input low-resolution image into the desired scale. The super-resolution is performed by the nearest neighbor based estimation of high-frequency patches based on the corresponding patches of input low-frequency image. Brandi et al. [7] propose an example-based approach for video super-resolution. They restore the high-frequency

Information of an interpolated block by searching in a database for a similar block, and by adding the high frequency of the chosen block to the interpolated one. They use the high frequency of key HR frames instead of the database to increase the quality of non-key restored frames. In [8], the authors address the problem of super-resolution from a single image using multi-scale tensor voting framework. They consider simultaneously all the three color channels to produce a multi-scale edge representation to guide the process of high-resolution color image reconstruction, which is subjected to the back projection constraint. The authors in [9] recover the super-resolution image through neighbor embedding algorithm. They employ histogram matching for selecting more reasonable training images having related contents. In [10] authors propose a neighbor embedding based super-resolution through edge detection and Feature Selection (NeedFS). They propose a combination of appropriate features for preserving edges as well as smoothing the color regions. The training patches are learned with different neighborhood sizes depending on edge detection. The authors in [11] propose modeling methodology for texture images. They

capture the features of texture using a set of filters which represents the marginal distribution of image and match the same in feature fusion to infer the solution. In this paper, we propose an approach to obtain super-resolution from a single image. First, we learn the high frequency content of the super-resolved image from the high-resolution training images in the data base and use the learnt image as a close approximation to the final solution. We solve this ill-posed problem using prior information in the form of marginal distribution. We apply different filters on the image and calculate the histograms. We assume that these histograms remain deviate from that of the close approximation. We show the result of our method on real images and compare it with the existing approaches.

.

II. DCT BASED APPROACH FOR CLOSE APPROXIMATION.

In this section, DCT based approach to learn high frequency details for the super-resolved for a decimation factor of 2 (q = 2) is described. Each set in the database consists of a pair of low resolution and high resolution image. The test image and LR training images are of size M × M pixels. Corresponding HR training images have size of 2M × 2M pixels. We first up sample the test image and all low resolution training images by factor of 2 and create images of size 2M X 2M pixels each. A standard interpolation technique can be used for the same. We divide each of the images, i.e. the up sampled test image, up sampled low resolution images and their high resolution versions, in blocks of size4 × 4. The motivation for dividing into 4X4 block is due to the theory of JPEG compression where an image is divided into 8X8 blocks in order to extract the redundancy in each block. However, in this case we are interested in learning the non aliased frequency components from the HR training images using the aliased test image and the aliased LR training images. This is done by taking the DCT on each of the block for all the images in the database as well as the test image. Fig.1.a) shows the DCT blocks of the up sampled test image whereas Fig.1. (b) Shows the DCT blocks of up sampled LR training images and HR training images. We learn DCT coefficients for each block in the test image from the corresponding blocks in the HR images in the database. It is reasonable to assume that when we interpolate the test image and the low resolution training images to obtain 2M × 2M pixels, the distortion is minimum in the lower frequencies. Hence we can learn those DCT coefficients that correspond to high frequencies (already aliased) and now distorted due to interpolation. We consider up sampled LR training images to find the best matching DCT coefficients for each of the blocks in the test image.

Let C_T (i , j), $1 \leq$ (i , j) $\leq$ = 4, be the DCT coefficient at location (i, j) in a 4 × 4 block of the test image. Similarly, let $C^{(m)}_{LR}$ (i , j) and $C^{(m)}_{HR}$ (I , j), m = 1,2,……,L, be the DCT coefficients at location(i, j) ,in the block at the same position in the mth up-sampled LR image and m^{th} HR image. Here L

is the number of the training sets in the data base. Now the best matching HR block for the considered low resolution image block (up-sampled) is obtained as

$$\hat{m} = argmin \sum_{i+j>Threshold}^{8} \left\| C^{(m)}_T(i,j) - C^{(m)}_{LR}(i,j) \right\|^2 .$$
(1)

Here, $\hat{m}(i, j)$ is the index for the training image which gives the minimum for the block. Those non aliased best matching HR image DCT coefficients are now copied in to the corresponding locations in the block of the up sampled test image. In effect, we learn non aliased DCT coefficients for the test image block from the set of LR-HR images. The coefficients that correspond to low frequencies are not altered. Thus at location (i , j)in a block, we have,

$$X_T(\iota, \varphi) = \begin{cases} C^{(\widehat{m})}_{HR}(i,j) & if\ (i,j) > Threshold \\ C_{T(i,j)} & else \end{cases}$$
(2)

This is repeated for every block in the test image. We conducted experiment with different Threshold values. We begin with Threshold =2 where all the coefficients except the DC coefficient are learned. We subsequently increased the threshold value and conducted the experiment. The best results were obtained when the Threshold was set to 4 that correspond to learning a total of 10 coefficients from the best matching HR image in the database. After learning the DCT coefficients for every block in the test image, we take inverse DCT transform to get high spatial resolution image and consider it as the close approximation to the HR image.

III. IMAGE FORMATION MODEL

In this work, we obtain super-resolution for an image from a single observation. The observed image Y
is of size M x M pixels. Let y represent the lexicographically ordered vector of size $M^2 × 1$, which contains the pixels from image Y and z be the super-resolved image. The observed images can be modeled as

$$y = Dz + n,$$
(3)

where D is the decimation matrix which takes care of aliasing. For an integer decimation factor of q, the decimation matrix D consists of q^2 non-zero elements along each row at appropriate locations. We estimate this decimation matrix from the initial estimate. The procedure for estimating the decimation matrix is described below. n is the i.i.d noise vector with zero mean and variance σ_n^2. It is of the size, $M^2 × 1$. The multivariate noise probability density is given by $2\pi\sigma_n^2$.

Our problem is to estimate z given y, which is an ill-posed inverse problem. It may be mentioned here that the observation captured is not blurred. In other words, we assume

identity matrix for blur. Generally, the decimation model to obtain the aliased pixel intensities from the high resolution pixels, for a decimation factor of q, has the form [12]

$$D = \frac{1}{q^2}\begin{pmatrix} 1\,1\,...1 & & 0 \\ & 1\,1\,...\,1 & \\ 0 & & 1\,1\,....1 \end{pmatrix} \qquad (4)$$

The decimation matrix in Eq. (4) indicates that a low resolution pixel intensity Y (i, j) is obtained by averaging the intensities of q^2 pixels corresponding to the same scene in the high resolution image and adding noise intensity n (i, j).
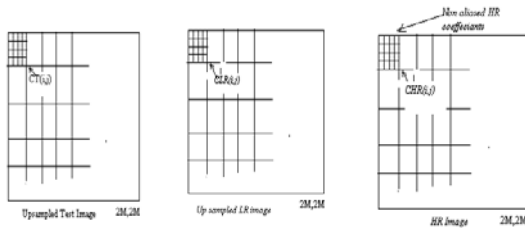
Training Set-1

•

•

•
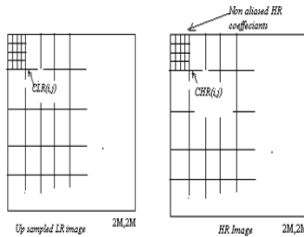
Training set L

Figure-1

IV TEXTURE MODELLING

Natural images consist of smooth regions, edges and texture areas. We regularize the solution using the texture preserving prior. We capture the features of texture by applying different filters to the image and compute histograms of the filtered images. These histograms estimate the marginal distribution of the image. These histograms are used as the features of the image. We use a filter bank that consists of two kinds of filters: Laplacian of Gaussian (LoG) filters and Gabor filters.

A. Filter Bank

The Gaussian filters play an important role due to its nice low pass frequency property. The two dimensional Gaussian function can be defined as

$$G(x,\,y\backslash x_0,y_0,\sigma_x,\sigma_y) = \frac{1}{2\pi\sigma_{x\sigma_y}} e^{-\left(\frac{x-x_0}{2\sigma^2}+\frac{y-y_0}{2\sigma 2}\right)} \qquad (5)$$

Here (x_0, y_0) are location parameters and $(\sigma_x \ \sigma_y)$ are scale parameters. The Laplacian of Gaussian (LoG) filter is a radially symmetric centers around Gaussian filter with $(x_0, y_0) = (0, 0)$ and $\sigma_x = \sigma_y = T$. Hence LoG filter can be represented by

$$F(x,\,y\backslash 0,0,\,T) = c\cdot(x^2+y^2-T^2)e^{-\frac{x^2+y^2}{T^2}} \qquad (6)$$

Here c is a constant and T scale parameter. We can choose different scales with $T = \frac{1}{\sqrt{2}}$, 1, 2,3, and so on. The Gabor filter with sinusoidal frequency 'ω 'and amplitude modulated by the Gaussian function can be represented by

$$F_w(x,y) = G(x,y\backslash 0,0,\sigma_x,\,\sigma_y)e^{-j\omega\theta} \qquad (7)$$

A simple case of Eq. (7) with both sine and cosine components can chosen as

$$G(x,y\backslash 0,\,0,T,\,\theta) = ce^{\frac{1}{2T^2}}\ (\ 4(xcos\theta+ysin\theta)^2\ + (-xsin\theta+ycos\theta)2 \qquad (8)$$

By varying frequency and rotating the filter in - y plane, we can obtain a bank of filters. We can choose different scales T=2, 4, 6, 8 and so on. Similarly, the orientation can be varied as $\theta = 0^0,30^0,60^0,90^0$ and so on.

B. Marginal Distribution Prior

As mentioned earlier, the histograms of the filtered images estimate the marginal distribution of the image. We use this marginal distribution as a prior. We obtain the close approximation Z_C of the HR image using discrete cosine transform based learning approach as described in section II and assume that the marginal distribution of the super-resolved image should match that of the close approximation Z_C. Let B be a bank of filters. We apply each of the filters in B to Z_C and obtain filtered images), where Z_C^α , where α = 1,. . . . . . ,|B|. We compute histogram $H_C^{(\alpha)}$ of $Z_C^{(\alpha)}$. Similarly, we apply each of the filter in B to the initial HR estimate and obtain filtered images Z^α ,where $\alpha = 1, 2, 3…. |B|$. We compute histogram H_C^α of Z_C^α . We define the marginal distribution prior term as,

$$C_H = \sum_{\alpha=1}^{|B|}|H_C^\alpha\ -\ H^\alpha| \qquad (9)$$

V. SUPER-RESOLVING THE IMAGE

The final cost function consisting of the data fitting term and marginal distribution prior term can be expressed as

$$\hat{Z} = \text{argument min}\left(\frac{\|y - Dz\|^2}{2\sigma_n^2} + \lambda \sum_{\alpha=1}^{|B|} |H_C^\alpha - H^\alpha| \right).$$
(10)

Where, λ is a suitable weight for the regularization term. The cost function consists of non-linear term it cannot be minimized using simple gradient descent optimization technique. We employ particle swarm optimization and avoid the computationally complex optimization methods like simulated annealing. Let S be the swarm. The swarm S is populated of images Zp, p =1……$|S|$ expanded using existing interpolation techniques such as bi-cubic interpolation, lanczose interpolation and learning based approaches. Each pixel in this swarm is a particle. The dimension of the search space for each image is
D = N × N. The i-th image of the swarm can be represented by a D-dimensional vector, Z =$(Zi1, Z_i2,…, Z_iD)^T$. The velocity of particles in this image can be represented by another D-dimensional vector

V_i = $(vi1; vi2,….., viD)^T$. The best previously visited position of the i-th image is denoted as

Pi = $(pi1, pi2;…., piD)^T$ .Defining ' g ' as the index of the best particle in the swarm, the swarm is manipulated according to the following two equations [13],

$$v_{id}^{n+1} = w v_{id}^n + c_1 r_1 (p_{id}^n - z_{id}^n)_+ c_2 r_2 (p_{gd}^n - z_{id}^n)$$
(11)

$$Z_{id}^{n+1} = Z_{id}^n + v_{ig}^n$$
(12)

where d =1, 2,….,D; i=1, 2,.….F; w is weighting function, r1 and r2 are random numbers uniformly distributed V1,V2,…are the iteration numbers, C1,C2,.. are cognitive and social parameter, respectively. The fitness function in our case is the cost function that has to be minimized.

VI. EXPERIMENTAL RESULTS

In this section, we present the results (shown in the fig.2, fig.3 and table-1) of the proposed method for the super-resolution. We compare the performance of the proposed method on the basis of quality of images. All the experiments were conducted on real images. Each observed image is of size 128 ×128 pixels. The super-resolved images are also of size 128 × 128. We used the quantitative measure Mean Square Error (MSE) for comparison of the results. The MSE used here is

$$\text{M.S.E} = \frac{\sum_{i,j} |f(i,j) - \widehat{f(i,j)}|^2}{\sum_{i,j} |f(i,j)|^2}$$

where f (i, j)is the original high resolution image and $\hat{f}$(i , j) is estimated super-resolution image.

(a)   HR Image

(b) Learnt Image

(c) PSO Optimized Image

Figure-2

(a) HR Image

(b) Learnt Image

(c)PSO Optimized Image
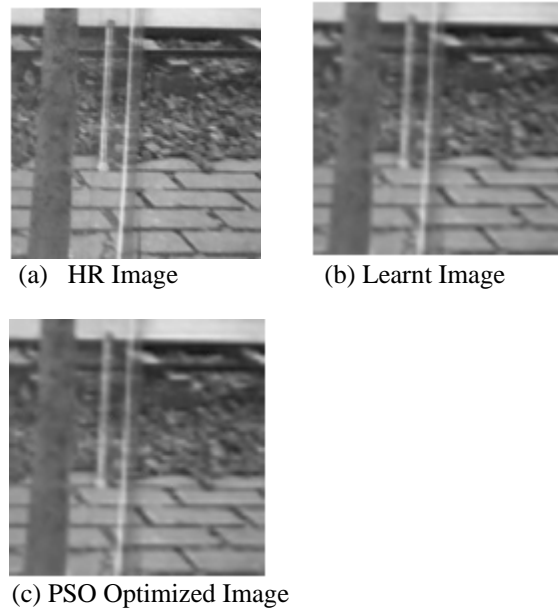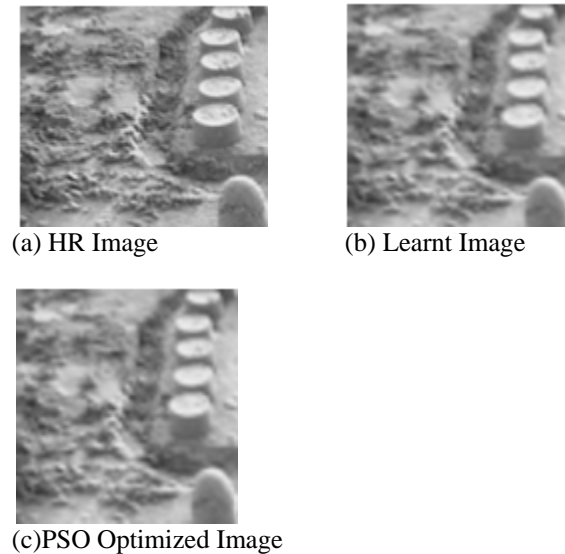
Figure-3

TABLE-1

Image Num	MMSE between HR and Learnt images	MMSE between HR and PSO images
1	0.02173679178	0.02154759509
2	0.01117524672	0.01107802761

VII.CONCLUSION.

We have presented a technique to obtain super-resolution for an image captured using a low cost camera. The high frequency content of the super-resolved image is learnt from a database of low resolution images and their high resolution versions. The suggested technique for learning the high frequency content of the super-resolved image yields close approximation to the solution. The LR observation is represented using linear model and marginal distribution is used as prior information for regularization. The cost function consisting of a data fitting term and a marginal distribution prior term is optimized using particle swarm optimization. The optimization process converges rapidly. It may be concluded that the proposed method yields better results considering both smoother regions as well as texture regions and greatly reduces the optimization time.

REFERENCES

[1]R.Y.Tsai and T.S.Huang, "Multiframe image resolution and registration" Advances in computer vision and image processing, pp.317-339, 1984.

[2] R.C. Hardle, K.J. Barnard, and E.E. Armstrong. "Joint MAP registration and high-resolution image estimation using a sequence of under sampled images", IEEE Trans.Image Process,vol.6.no12,pp.1621-1633,Dec.1997.

[3] D. Rajan and S. Chaudhuri, "Generation of super-resolution images from blurred observation using an MRF model,".ath.Imag.ision, Vol.16, pp,5-15, 2002.

[4] S.Chaudhuri, Super-resolution imaging, S.Chaudhuri, Ed.kluwer, 2001.

[5] C.V.Jiji, M.V.Joshi, and S. Chaudhuri, "Single frame image super-resolution using learned wavelet coefficients. "International Journal of Imaging Systems and Technology, Vol.14,no.3,pp.105-112,2004.

[6] W.Freeman, T.Jones, E.Pasztor, "Example based super-resolution," IEEE Computer GRAPHICS AND asPPLICATIONS, Vol.22,no.2,pp.56-65,2002.

[7]F.Brandi, R.de Queiroz, and D.Mukerjee, "super-resolution of video using key frames,"IEEE International Symposium on circuits and systems,pp.1608-1611,2008.

[8]Y.W.Tai, W.S.tongand C.K.Tang,"Perceptually inspired and edge-directed color image super-resolution,"IEEE Computer Society Conference on Computer Vision and Pattern recognition, vol.2,pp.1948-1955,2006.

[9] T.Chan and J.Zhang,"An improved super-resolution with manifold learning and histogram matching,"Proc.IAPRInternational Conference on Biometric,pp.756-762,2006.

[10] T.Chan,J.Zhang, J.Pu, and H.Huang,"Neighbot Embedding based super-resolution algorithm through edge detection and feature selection,"Pattern Recognition Letters, Vol.30,no.5,pp,494-502,2009.

[11] W.Y.Zhu, S.C. and D.Mumford,"Filters,random fields and aximum entropy(FRAME): Towards unified theory for texture modeling,"International Journal of computer Vision, vol.27,no.2,pp.107-126, 1998.

[12] R.R.Schultz and R.L.Stevenson,"A Bayseian approach to image expansion for improved definition," IEEE Trans.Image.Process,vol.3, no.3,pp233-242, May 1994.

[13] M.Vrahatis and K.Parsopoulous, " Natural Computing," Kluwer , 2002.

AUTHORS PROFILE

S.Ravishankar did his M.Tech fro National Institute of Technology,Surathkal, India. Currently working faculty in the deprtment of ECE at Amrita Vishwavidyapeetham University,Bangalore campus, India. His area of research includes Image super resolution,Digital Image watermarking, Machine learning and VLSI implementation for image processing algorithms.

K.V.V.Murthy B.Sc; B.E;M.Tech;PhD; Currently working as distinguished professor in the department of Electrical Engineering Indian Institute of Technology, Gandhinagar,Gujarath, India. His research area includes Signal processing,Image ocessing,Wavewlets, Electrical Network theory, Electrical filters ,CAD for VLSI design, and Engineering education.

A Survey on WiMAX

Mohsen Gerami

The Faculty of Applied Science of Post and Communications
Danesh Blv, Jenah Ave, Azadi Sqr, Tehran, Iran.
Postal code: 1391637111
e-mail: artimes0@hotmail.com

Abstract—**This paper describes an overview of WiMAX. The paper outlines fundamental architectural components for WiMAX and explains WiMAX Security Issues. Furthermore various 802.16 standards, IEEE 802.16 protocol architecture and WiMAX Market will be discussed.**

Keywords: WiMAX; IEEE 802.16; Security; Protocol; Market;

I. INTRODUCTION

WiMAX, meaning Worldwide Interoperability for Microwave Access, is a telecommunications technology that provides wireless transmission of data using a variety of transmission modes, from point-to-multipoint links to portable and fully mobile internet access. The technology provides up to 10 Mbps broadband speed without the need for cables. The technology is based on the IEEE 802.16 standard (also called Broadband Wireless Access). The name "WiMAX" was created by the WiMAX Forum, which was formed in June 2001 to promote conformity and interoperability of the standard. The forum describes WiMAX as "a standards-based technology enabling the delivery of last mile wireless broadband access as an alternative to cable and DSL" [1].

As compared to a wireless technology like Wi-Fi, WiMAX is more immune to interference, allows more efficient use of bandwidth and is intended to allow higher data rates over longer distances. Because it operates on licensed spectrum, in addition to unlicensed frequencies, WiMAX provides a regulated environment and viable economic model for wireless carriers. These benefits, coupled with the technology's global support (e.g., ongoing worldwide deployments, spectrum allocation and standardization), make it the popular choice for quick and cost-effective delivery of super-fast broadband wireless access to underserved areas around the world [2].

WiMAX is cheaper than wired DSL because it does not require placing wires around the area to be covered, which represents an enormous investment for the provider. Not requiring this investment opens the door to many service providers who can start retailing out wireless broadband with low capital, thereby causing prices to drop due to competition .

As with any wireless technology, the requirements for WiMAX are basically a transmitter and a receiver. The transmitter is a WiMAX tower, much like a GSM tower. it is the part of the service provider's facilities. One tower, also called a base station, can provide coverage to an area within a radius of around 50 km. On the other side, in order to receive the WiMAX waves, you need a receiver for WiMAX for connecting your computer or device.

WiMAX has a range of around 50 km in a circle. Terrain, weather and buildings affect this range and this often results in many people not receiving signals good enough for a proper connection. Orientation is also an issue, and some people have to choose to place their WiMAX modems near windows and turned in certain specific directions for good reception.

A WiMAX connection is normally non-line-of-sight, which means that the transmitter and the receiver need not have a clear line between them. But a line-of-sight version exists, where performance and stability is much better, since this does away with problems associated with terrain and buildings [3].

II. WIMAX FUNDAMENTAL ARCHITECTURAL COMPONENTS

WiMAX has four fundamental architectural components:

Base Station (BS). The BS is the node that logically connects wireless subscriber devices to operator networks. The BS maintains communications with subscriber devices and governs access to the operator networks. A BS consists of the infrastructure elements necessary to enable wireless communications, i.e., antennas, transceivers, and other electromagnetic wave transmitting equipment. BSs are typically fixed nodes, but they may also be used as part of mobile solutions—for example, a BS may be affixed to a vehicle to provide communications for nearby WiMAX devices. A BS also serves as a Master Relay-Base Station in the multi-hop relay topology.

Subscriber Station (SS). The SS is a fixed wireless node. An SS typically communicates only with BSs, except for multi-hop relay network operations. SSs are available in both outdoor and indoor models.

Mobile Subscriber (MS). Defined in IEEE 802.16e-2005, MSs are wireless nodes that work at vehicular speeds and support enhanced power management modes of operation. MS devices are typically small and self-powered, e.g., laptops, cellular phones, and other portable electronic devices.

Relay Station (RS). Defined in IEEE 802.16j-2009, RSs are SSs configured to forward traffic to other RSs, SSs, or MSs in a multi-hop Security Zone [4].
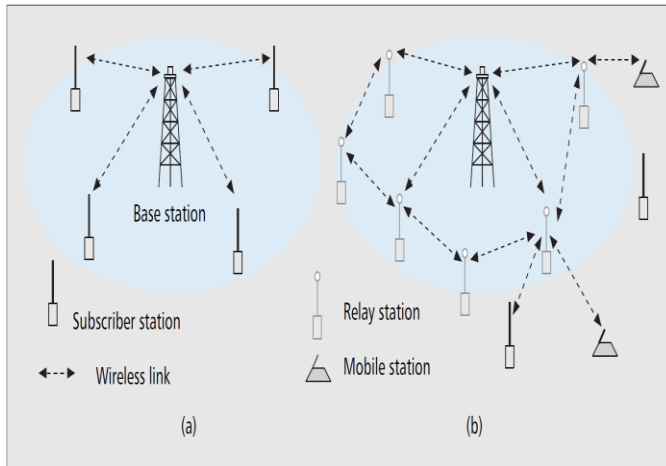
Figure 1.   WiMAX network architectures: (a) PMP mode; (b) mesh mode [5].

WiMAX devices communicate using two message types: management messages and data messages. *Data messages* transport data across the WiMAX network. *Management messages* are used to maintain communications between an SS/MS and BS, i.e., establishing communication parameters, exchanging security settings, and performing system registration events (initial network entry, handoffs, etc.)

IEEE 802.16 defines frequency bands for WiMAX operations based on signal propagation type. In one type, WiMAX employs a radio frequency (RF) beam to propagate signals between nodes. Propagation over this beam is highly sensitive to RF obstacles, so an unobstructed view between nodes is needed. This type of signal propagation, called *line-of-sight (LOS)*, is limited to fixed operations and uses the 10–66 gigahertz (GHz) frequency range. The other type of signal propagation is called *non-line-of-sight (NLOS)*. NLOS employs advanced RF modulation techniques to compensate for RF signal changes caused by obstacles that would prevent LOS communications. NLOS can be used for both fixed WiMAX operations (in the 2–11 GHz range) and mobile operations (in the 2–6 GHz range). NLOS signal propagation is more commonly employed than LOS because of obstacles that interfere with LOS communications and because of strict regulations for frequency licensing and antenna deployment in many environments that hinder the feasibility of using LOS [4].

III.   IEEE 802.16

The IEEE developed the 802.16 in its first version to address line of sight (LOS) access at spectrum ranges from 10 GHz to 66 GHz. The technology has evolved through several updates to the standard such as 802.16a, 802.16c, the Fixed WiMAX 802.16d (802.16-2004) specification and lastly the mobile 802.16e set that are currently commercially available. The upcoming 802.16m is still a ways away from ratification. The first update added support for 2 GHz through 11 GHz spectrum with NLOS capability. Each update added additional functionality or expanded the reach of the standard.

For example, the 802.16c revision added support for spectrum ranges both licensed and unlicensed from 2 GHz to 10 GHz. It

also improved quality of service (QOS) and certain improvements in the media access control (MAC) layer along with adding support for the HiperMAN European standard. The number of supported physical (PHY) layers was increased. Transport mediums such as IP, Ethernet and asynchronous transfer mode (ATM) were added.

At its core, the technology is intended to take a number of best of breed proprietary enhancements that had been made by vendors using the 802.11 standard and combine them together in a very marketable and standardized WiMAX product.

For example, older broadband wireless technology such as the Wi-Fi or 802.11b system utilized carrier sense multiple access with collision detection (CSMA/CD) crosstalk methods for base stations and customer premise equipment (CPE) to talk to one another. Basically, this meant that each radio was constantly talking and creating inefficient overhead. It also resulted, especially at times of high traffic, in increased packet collisions and retransmissions, further exacerbating the problem. Some of the proprietary MAC systems built later utilized the base station to define when the CPE would be polled in order to eliminate this problem. In the way of a permanent cure the 802.16 protocol supports multiple methods of polling that a vendor can choose to use. Some of these include piggybacking polling requests within overhead traffic, group polling or dynamic co-opting of bandwidth from another unit by the CPE. The key is that the radios will be interchangeable based on the Forum's initial product profile as well as more efficient [6].

A.   The various 802.16 standards

802.16a: Licensed Frequency 2 GHz to 11 GHz. The Working IEEE 802.16a operates at the MAC and PHY specification and specifies the transfer of non-visual connections (NLOS). Frequencies are important for the 3.5 GHz and 5.8 GHz licensed for royalty-free applications. The data is at a channel width of 20 MHz 75 Mbit / s. 802.16a is replaced by 802.16-2004.

Specifications of 802.16

802.16b: Licensed Exempt Frequencies, with a focus on the frequency band of between 5 GHz and 6 GHz. This group also runs under the name Wireless HUMAN (High Speed Unlicensed MAN).

802.16c: Profiles of transmission frequencies in the frequency range from 10 GHz to 66 GHz. The channel width is in the U.S. 25 MHz, 28 MHz in Europe. 802.16c is replaced by 802.16-2004.

802.16d: Profiles of transmission frequencies in the frequency range of 2 GHz to 66 GHz. Replaced by 802.16-2004. This standard provides visual and non-visual connections in the range of 2 GHz to 66 GHz.

802.16e-2005: Mobile Wireless MAN (WMAN). This working group defines a mobile access in the context of IEEE 802.16. Here are ranges of more than 10 Mbps in cells in the range of several kilometers and speeds exceeding 100 kph investigated. In addition, 16e-clients between different radio

cells can switch, known as roaming. 802.16e is in conjunction with DSRC an interesting alternative for telematic and safety services in the automotive technology.

802.16f: MIB management for access networks.

802.16g: Definition of Management Plane.

802.16h: Coexistence of Networks. This Working Group deals with the problems of coexistence of different radio technologies in unlicensed bands transmission.

802.16i: Mobile One Plane Information

802.16j: bridging alternative to 802.11k. This involves Equipment for a mobile relay, which has several communications partner stations can connect.

802.16k: Bridging

802.16m: 802.16m The group is working on the high-speed transmission with up to 1 Gbit / s.

802.16-1: Air Interface for 10 GHz to 66 GHz.

802.16.2: Coexistence of Broadband Wireless Access Systems. This Working Group deals with the coexistence of existing systems. Replaced by 802.16.2-2004.

802.16.2-2004: Combines standards 802.16, 802.16a, 802.16c and 802.16d in a standard and regulate the coexistence of wireless broadband systems in the range of 10 GHz to 66 GHz.

802.16.2a: Recommended Practice for Coexistence of Fixed Broadband Wireless Access Systems. This group is the coexistence of PMP systems between 2 GHz and 11 GHz redefine.

802.16.3: Air Interface for Fixed Broadband Wireless Access Systems operating below 11 GHz. In this group are the unlicensed bands, such as the ISM band, the Personal Communications Services (PCS), and MMDS Unii for the use of a high-speed access MAN investigated [7].

The following table provides a summary of the IEEE 802.16 family of standards [8].

TABLE I.   SUMMARY OF THE IEEE 802.16 FAMILY OF STANDARDS

Standard	802.16	802.16a/802.16REVd	802.16e
Spectrum	10 to 66 GHz	< 11 GHz	< 6 GHz
Channel Conditions	Line-of-Sight only	None-Line-of-Sight	Non-Line-of-Sight
Speed (bit rate)	32 to 134 Mbps	75 Mbps max, 20-MHz channelization	15 Mbps max, 5-MHz channelization
Modulation	QPSK 16QAM 64QAM	OFDM 256 subcarrier QPSK 16QAM 64QAM	same as 802.16a
Mobility	Fixed	Fixed	Pedestrian mobility, regional roaming
Channel Bandwidths	20, 25 and 28 MHz	Selectable between 1.25 and 20 MHz	same as 802.16a with sub-channels
Typical Cell Radius	1 – 3 miles	3-5 miles (up to 30 miles, depending on tower height, antenna gain and transmit power)	1-3 miles

B. IEEE 802.16 protocol architecture

The IEEE 802.16 protocol architecture is structured into two main layers: the Medium Access Control (MAC) layer and the Physical (PHY) layer, as described in the following table [9]:
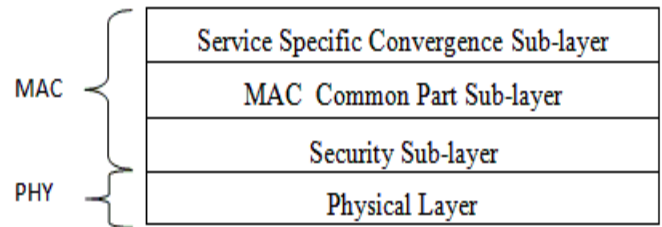
Figure 2.   The IEEE 802.16 Protocol structure

MAC layer consists of three sub-layers. The first sub-layer is the Service Specific Convergence Sub-layer (CS), which maps higher level data services to MAC layer service flow and connections [10]. The second sub-layer is Common Part Sub-layer (CPS), which is the core of the standard and is tightly integrated with the security sub-layer. This layer defines the rules and mechanisms for system access, bandwidth allocation and connection management. The MAC protocol data units are constructed in this sub-layer. The last sub-layer of MAC layer is the Security Sub-layer which lies between the MAC CPS and the PHY layer, addressing the authentication, key establishment and exchange, encryption and decryption of data exchanged between MAC and PHY layers.

The PHY layer provides a two-way mapping between MAC protocol data units and the PHY layer frames received and transmitted through coding and modulation of radio frequency signals [8].

IV. WiMAX Security

Realizing the sticking point that security has been in the widespread adoption of broadband wireless service, the IEEE and the Forum both determined to define a robust security environment. WiMAX security supports two quality encryptions standards, that of the DES3 and AES, which is considered leading edge. The standard defines a dedicated security processor on board the base station for starters. There are also minimum encryption requirements for the traffic and for end to end authentication---the latter of which is adapted from the data-over-cable service interface specification (DOCSIS) BPI+ security protocol.

Basically, all traffic on a WiMAX network must be encrypted using Counter Mode with Cipher Block Chaining Message Authentication Code Protocol (CCMP) which uses AES for transmission security and data integrity authentication.

The end-to-end authentication the PKM-EAP (Extensible Authentication Protocol) methodology is used which relies on the TLS standard of public key encryption. At least one chip company designed processors to support this standard of onboard security processor [11].

A. WiMAX security solutions

By adopting the best technologies available today, the WiMAX, based on the IEEE 802.16e standard, provides strong support for authentication, key management, encryption and decryption, control and management of plain text protection and security protocol optimization. In WiMAX, most of security issues are addressed and handled in the MAC security sub-layer as described in the following figure:
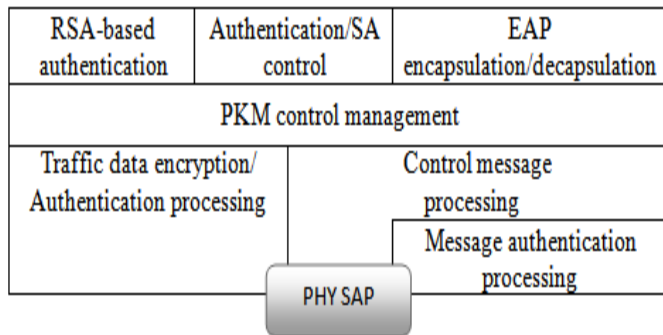
Figure 3. MAC Security sub-layer .

Source: IEEE Std. 802.16e 2006.

Two main entities in WiMAX, including Base Station (BS) and Subscriber Station (SS), are protected by the following WiMAX security features[8]:

Security associations: A security association (SA) is a set of security information parameters that a BS and one or more of its client SSs share in order to support secure communications. Data SA has a 16bit SA identifier, a Cipher (DES in CBC mode) to protect the data during transmission over the channel and two traffic encryption keys (TEKs) to encrypt data: one is the current operational key and the other is TEK [12]. When the current key expires, TEK a 2bit key

identifiers is used. A 64bit initialization vector (IV) is used for each TEK [13].

Public key infrastructure (PKI): The WiMAX standard uses the Privacy and Key Management Protocol for securely transferring keying material between the base station and the mobile station. The privacy key management (PKM) protocol is responsible for privacy, key management, and authorizing an SS to the BS. The initial draft for WiMAX mandates the use of PKMv1 [14], which is a one-way authentication method. PKMv1 requires only the SS to authenticate itself to the BS, which poses a risk for a Man-in-the-middle (MITM) attack. To overcome this issue, PKMv2 was proposed (later adopted by 802.16e), which uses a mutual (two-way) authentication protocol [15]. Here, both the SS and the BS are required to authorize and authenticate each other. PKMv2 is preventing from the following [16]: BS and SS impersonations, MITM attack and Key exchange issue.

PKMv2 supports the use of the Rivest-Shamir-Adlerman (RSA) public key cryptography exchange. The RSA public key exchange requires that the mobile station establish identity using either a manufacturer-issued X.509 digital certificate or an operator-issued credential such as a subscriber identity module (SIM) card. The X.509 digital certificate contains the mobile station's Public-Key (PK) and its MAC address. The mobile station transfers the X.509 digital certificate to the WiMAX network, which then forwards the certificate to a certificate authority. The certificate authority validates the certificate, thus validating the user identity.
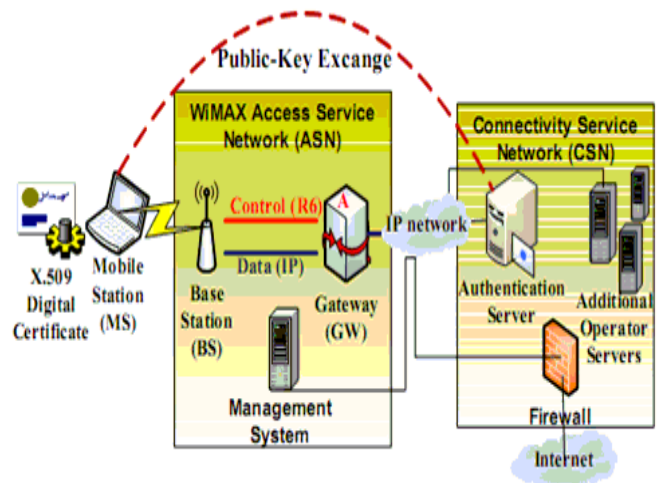
Figure 4. Public Key Infrastructure [13].

Once the user identity is validated, the WiMAX network uses the public key to create the authorization key, and sends the authorization key to the mobile station. The mobile station and the base station use the authorization key to derive an identical encryption key that is used with the advanced encryption standard (AES) algorithm [13].

Authentication: Authentication is the process of validating a user identity and often includes validating which services a user may access. The authentication process typically involves a supplicant (that resides in the mobile station), an

authenticator (that may reside in the base station or a gateway), and an authentication server [13].
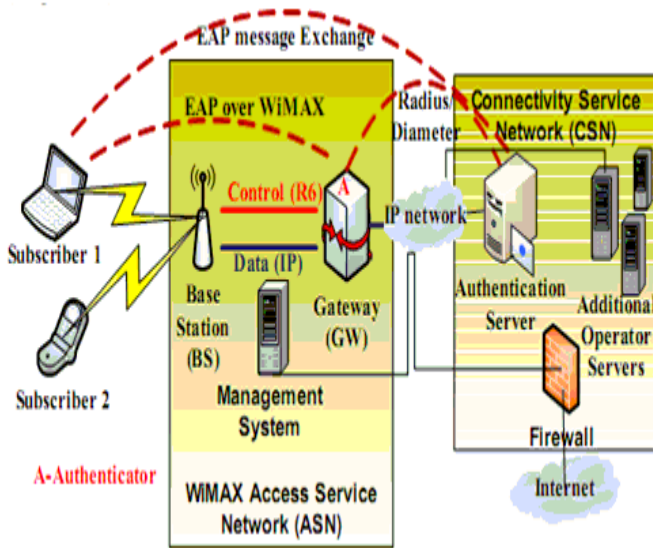
Figure 5.   EAP-based authentication [13].

WiMAX uses the Extensible Authentication Protocol (EAP) to perform user authentication and access control. EAP is actually an authentication framework that requires the use of "EAP methods" to perform the actual work of authentication. The network operator may choose an EAP method such as EAP-TLS (Transport Layer Security), or EAP-TTLS MS-CHAP v2 (Tunneled TLS with Microsoft Challenge-Handshake Authentication Protocol version 2). The messages defined by the EAP method are sent from the mobile station to an authenticator. The authenticator then forwards the messages to the authentication server using either the RADIUS or DIAMETER protocols [17].

Data privacy and integrity: WiMAX uses the AES to produce ciphertext. AES takes an encryption key and a counter as input to produce a bitstream. The bitstream is then XORed with the plaintext to produce the cipher text. AES algorithm is the recommendation of 802.16e security sub-layer, since it can perform stronger protection from theft of service and data across broadband wireless mobile network. Besides CCM-Mode and ECB-Mode AES algorithm supported in 802.16-2004, 802.16e supports three more AES algorithms: CBC-Mode AES, CTR-Mode AES and AES-Key-Wrap[13].

B.   WiMAX threats

Despite good intentions for WiMAX security, there are several potential attacks open to adversaries, including:

Rogue Base Stations

DoS Attacks

Man-in-the-Middle Attacks

Network manipulation with spoofed management frames

The real test of WiMAX security will come when providers begin wide-scale network deployments, and researchers and attackers have access to commodity CPE equipment. Other attacks including WiMAX protocol fuzzing may enable attackers to further manipulate BSs or SSs. Until then, the security of WiMAX is limited to speculation [18].

V.   GLOBAL WIMAX MARKET

World Interoperability for Microwave Access or WiMAX, has been gaining a lot of attention as a wireless broadband alternative, as it provides reliable, secure and high quality broadband access for mobile Internet users. The technology supports bandwidth-heavy applications and User Generated Content (UGC) services that customers want. WiMAX promises a better-performing, less-expensive alternative to many technologies (like DSL, Wi-Fi) that are already available in the market.

According to new research report "Global WiMAX Market Analysis", WiMAX has tremendous potential to offer global standardized broadband wireless platform. Many countries across the globe will adopt WiMAX to facilitate rapid economic development. Moreover, the move to WiMAX, a technology that is ready for deployment now, will be preferable to waiting for alternative technologies that may not be available for three or more years. As a result, the number of WiMAX users is forecast to grow over 87% between 2010 and 2012.

The research reveals that, by 2012 the Asia-Pacific region will lead the number of global WiMAX users accounting for over 45% of the total user base, followed by North America and Europe. Major growth is expected in Asia-Pacific and MEA as these countries are deploying the technology more rapidly. Moreover, government support and operators' initiatives to provide the region with faster Internet access in remote areas is also fostering growth into the WiMAX market [19].

The WiMAX market is coming out of the recession period strongly, posting three consecutive quarters of revenue growth for 802.16e equipment and devices. With Clearwire in the U.S. announcing strong quarterly results, Yota in Russia expanding rapidly, and others such as UQ in Japan being aggressive, the WiMAX business model seems to be working. Though we are still in the early days, WiMAX is proving to be a good fit in a range of broadband segments in developed as well as developing markets [20].

WIMAX MARKET HIGHLIGHTS

•   Worldwide vendor revenue from 802.16d and 802.16e WiMAX network equipment and devices hit $1.08 billion in 2009, down 19% from 2008, as the market suffered the effects of the recession

•   However, 4Q09 was the third consecutive quarter of WiMAX equipment and device revenue growth, up 3% from 3Q09

o   Quarterly revenue levels remain short of the pre-recession market highs of over $300 million seen in early 2008

• The WiMAX market is showing positive signs of steady growth from this year onward, with major rollouts underway in USA, Japan, Russia, and India

• Starting in 2011-2012, 802.16m WiMAX products are expected to be tested, certified, and commercially available, offering speeds comparable to LTE

• For the combined WiMAX equipment and device market, Motorola took the #1 spot in 2009, with 17% of worldwide revenue, just ahead of Alvarion

• Huawei showed the biggest growth in WiMAX equipment and device market share in 2009

• The number of WiMAX subscribers jumped 75% in 2009 to 6.8 million worldwide [21].

Figure 6.   WiMAX Market Forecast.

VI.   CONCLUSION

WiMAX allows operators to present their subscribers true broadband connectivity in fully mobile, all-IP networks. The IEEE 802.16e standard has changed several security mechanisms and need more research on its securities vulnerabilities. WiMAX is a very promising technology for delivery of fully mobile personal broadband services. WiMAX market presents enormous business opportunities. WiMAX can be deployed to drive new revenue streams on much shorter timelines and at much lower capex than FTTx, xDSL, or cable modem alternatives. WiMAX is an opportunity.

REFERENCES

[1]   WiMax Forum - Technology. http://www.wimaxforum.org/technology/. Retrieved 2008-07-22.

[2]   http://www.agilent.com/about/newsroom/tmnews/background/wimax/. Retrieved 2009-11-18.

[3]   Nadeem   Unuth, http://voip.about.com/od/mobilevoip/a/UsingWiMAXTechnology.htm. Retrieved 2009-10-12.

[4]   Karen Scarfone, Cyrus Tibbs, Matthew Sexton, 2009,Guide to Security for WiMAX Technologies, US National Institute of Standards and Technology-Special Publication 800-127(Draft), 46 pages (Sep. 2009)

[5]   David Johnston & Jesse Walker,2009, Overview of IEEE 802.16 security

[6]   http://slingbroadband.com/wimax/category/wimax-faq//   .   Retrieved 2008-11-28.

[7]   http://www.wifinotes.com/wimax/IEEE-802.16.html

[8]   8- Trung Nguyen, 2009, A survey of WiMAX security threats, http://www.cse.wustl.edu/~jain/cse571-09/ftp/wimax2/index.html

[9]   http://www.cse.wustl.edu/~jain/cse574-08/

[10] Department University of Bridgeport, Bridgeport, CT. http://www.asee.org/activities/organizations/zones/proceedings/zone1/20 08/Professional/ASEE12008_0022_paper.pdf

[11] http://slingbroadband.com/wimax/category/wimax-faq//   .   Retrieved 2008-11-28.

[12] J. Hasan, 2006, Security Issues of IEEE 802.16 (WiMAX), School of computer and Information Science, Edith Cowan University, Australia, 2006.

[13] Mitko Bogdanoski, Pero Latkoski, Aleksandar Risteski, Borislav Popovski,,2008,IEEE 802.16Security Issues: A Survey, Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius   University,   Skopje, Macedonia.,http://2008.telfor.rs/files/radovi/02_32.pdf

[14] D. Johnston and J. Walker, 2004, Overview of IEEE 802.16 Security, IEEE Security & Privacy, magazine May/June 2004.

[15] S. Adibi, G. B. Agnew,T. Tofigh, 2008,End-to-End (E2E) Security Approach in WiMAX: Security Technical Overview for Corporate Multimedia Applications, 747-758, Handbook of Research on Wireless Security (2 Volumes) Edited By: Yan Zhang, Jun Zheng, Miao Ma, 2008.

[16] S.Adibi, G. B. Agnew, 2008, End-to-End Security Comparisons Between IEEE 802.16e and 3G Technologies, 364 - 378, Handbook of Research on Wireless Security (2 Volumes) Edited By: Yan Zhang, Jun Zheng, Miao Ma, 2008.

[17] S. Adibi, G. B. Agnew, 2008, Extensible Authentication (EAP) Protocol Integrations in the Next Generation Cellular Networks, 776-789, Handbook of Research on Wireless Security (2 Volumes) Edited By: Yan Zhang, Jun Zheng, Miao Ma, 2008.

[18] Joshua   Wright, http://www.computerworld.com.au/article/170510/wimax_security_issu es/?fp=16&fpid=1, Network World

[19] Global   WiMAX   Market   Analysis,   2009, http://www.bharatbook.com/Market-Research-Reports/Global-WiMAX-Market-Analysis.html

[20] Webb Richard, 2010, London, United Kingdom, March 1, 2010— Infonetics Research

[21] WiMAX Equipment, Devices, and Subscribers market share and forecast report,2010, www.infonetics.com

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA

Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia

Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA

Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway

Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India

Dr. Amogh Kavimandan, The Mathworks Inc., USA

Dr. Ramasamy Mariappan, Vinayaka Missions University, India

Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China

Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA

Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico

Dr. Neeraj Kumar, SMVD University, Katra (J&K), India

Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania

Dr. Junjie Peng, Shanghai University, P. R. China

Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia

Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India

Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain

Prof. Dr.C.Suresh Gnana Dhas, Anna University, India

Mrs Li Fang, Nanyang Technological University, Singapore

Prof. Pijush Biswas, RCC Institute of Information Technology, India

Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia

Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India

Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand

Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India

Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia

Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India

Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India

Mr. P. Vasant, University Technology Petronas, Malaysia

Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea

Mr. Praveen Ranjan Srivastava, BITS PILANI, India

Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong

Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia

Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan

Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria

Dr. Riktesh Srivastava, Skyline University, UAE

Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia

Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt

 and Department of Computer science, Taif University, Saudi Arabia

Mr. Santosh K. Pandey, Department of Information Technology, The Institute of Chartered Accountants of India

Dr. P. Vasant, Power Control Optimization, Malaysia

Dr. Petr Ivankov, Automatika - S, Russian Federation

Dr. Utkarsh Seetha, Data Infosys Limited, India

Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal

Dr. (Mrs) Padmavathi Ganapathi, Avinashilingam University for Women, Coimbatore

Assist. Prof. A. Neela madheswari, Anna university, India

Prof. Ganesan Ramachandra Rao, PSG College of Arts and Science, India

Mr. Kamanashis Biswas, Daffodil International University, Bangladesh

Dr. Atul Gonsai, Saurashtra University, Gujarat, India

Mr. Angkoon Phinyomark, Prince of Songkla University, Thailand

Mrs. G. Nalini Priya, Anna University, Chennai

Dr. P. Subashini, Avinashilingam University for Women, India

Assoc. Prof. Vijay Kumar Chakka, Dhirubhai Ambani IICT, Gandhinagar ,Gujarat

Mr Jitendra Agrawal, : Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal

Mr. Vishal Goyal, Department of Computer Science, Punjabi University, India

Dr. R. Baskaran, Department of Computer Science and Engineering, Anna University, Chennai

Assist. Prof, Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India

Dr. Jamal Ahmad Dargham, School of Engineering and Information Technology, Universiti Malaysia Sabah

Mr. Nitin Bhatia, DAV College, India

Dr. Dhavachelvan Ponnurangam, Pondicherry Central University, India

Dr. Mohd Faizal Abdollah, University of Technical Malaysia, Malaysia

Assist. Prof. Sonal Chawla, Panjab University, India

Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India

Mr. Arash Habibi Lashkari, University of Malaya (UM), Malaysia

Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia

Professor Dr. Sabu M. Thampi, .B.S Institute of Technology for Women, Kerala University, India

Mr. Noor Muhammed Nayeem, Université Lumière Lyon 2, 69007 Lyon, France

Dr. Himanshu Aggarwal, Department of Computer Engineering, Punjabi University, India

Prof R. Naidoo, Dept of Mathematics/Center for Advanced Computer Modelling, Durban University of Technology, Durban,South Africa

Prof. Mydhili K Nair, M S Ramaiah Institute of Technology(M.S.R.I.T), Affliliated to Visweswaraiah Technological University, Bangalore, India

M. Prabu, Adhiyamaan College of Engineering/Anna University, India

Mr. Swakkhar Shatabda, Department of Computer Science and Engineering, United International University, Bangladesh

Dr. Abdur Rashid Khan, ICIT, Gomal University, Dera Ismail Khan, Pakistan

Mr. H. Abdul Shabeer, I-Nautix Technologies,Chennai, India

Dr. M. Aramudhan, Perunthalaivar Kamarajar Institute of Engineering and Technology, India

International Journal Computer Science and Information Security, now at its sixth edition, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2010 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA,   Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .